



REPUBLIC OF TÜRKİYE
ALTINBAŞ UNIVERSITY
Institute of Graduate Studies
Information Technologies

**PREDICTING USER CLICKS ON ONLINE
ADVERTISEMENTS USING MACHINE
LEARNING**

Ashraf Farhan Hatem AL-KHAFAJI

Master's Thesis

Supervisor

Asst. Prof. Dr. Oğuz KARAN

Istanbul, 2023

PREDICTING USER CLICKS ON ONLINE ADVERTISEMENTS USING MACHINE LEARNING

Ashraf Farhan Hatem AL-KHAFAJI

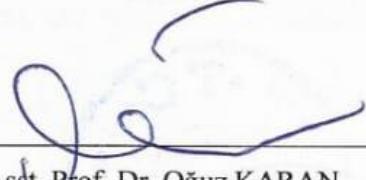
Information Technologies

Master's Thesis

ALTINBAŞ UNIVERSITY

2023

This thesis title PREDICTING USER CLICKS ON ONLINE ADVERTISEMENTS USING MACHINE LEARNING prepared by ASHRAF FARHAN HATEM AL-KHAFAJI and submitted on 29/12/2023 has been **accepted unanimously** for the degree of Master of Science in Information Technologies.


Asst. Prof. Dr. Oğuz KARAN

Supervisor

Thesis Defense Committee Members:

Asst. Prof. Dr. Oğuz KARAN Department of Software
Engineering,
Altınbaş University

Asst. Prof. Dr. Abdullahi Abdu
IBRAHIM Department of Computer
Engineering,
Altınbaş University


Asst. Prof. Dr. Zeynep ALTAN Department of Software
Engineering,
Beykent University

I hereby declare that this thesis meets all format and submission requirements of a Master's thesis.

Submission data of the thesis to Institute of Graduate Studies: ____/____/____

ADİ GEÇEN KURUM ÜLKEMİZİN
YÜKSEK ÖĞRENİM
KURUMLARINDANDIR.

01 - 02 - 2024


Emrah KOÇ
Şube Müdürü

REPUBLIC OF TURKEY
MINISTER OF FOREIGN AFFAIRS
Directorate General for Consular Affairs
This is to certify that the signature and
MILLİ EĞİTİM BAKANLIĞINA:
THIS CERTIFICATION DOES NOT COVER THE CONTENT
AND THE AUTHENTICITY OF THIS DOCUMENT
Number: 51467 01 Şubat 2024



Mine ÜNAL
İkinci Katip
KOPR

سفارة جمهورية العراق
المحققة الثقافية - انقره
تصديقات
التاريخ ٢٠٢٤ / ٢ / ٠١
Embassy of Iraq -
Cultural Attache - Ankara
رقم الوصل: ١٤٥٢١-
Fiş Numarası:

I hereby declare that all information/data presented in this graduation project has been obtained in full accordance with academic rules and ethical conduct. I also declare all unoriginal materials and conclusions have been cited in the text and all references mentioned in the Reference List have been cited in the text, and vice versa as required by the abovementioned rules and conduct.

Ashraf Farhan Hatem AL-KHAFAJI

Signature

ABSTRACT

PREDICTING USER CLICKS ON ONLINE ADVERTISEMENTS USING MACHINE LEARNING

AL-KHAFAJI, Ashraf Farhan Hatem

M.Sc., Information Technologies, Altınbaş University,

Supervisor: Asst. Prof. Dr. Oğuz KARAN

Date: 12/2023

Pages: 80

This thesis explores the detection of user behavior within online advertising to better understand user preferences and refine ad strategies. As online ads are instrumental in reaching broad audiences, targeted approaches are vital for enhancing campaign effectiveness. The research employs data analysis, preprocessing, and machine learning (ML) techniques on a dataset detailing user behaviors like browsing history, ad clicks, and demographics. ML methods, including random forest, GB, and LR, are utilized alongside XAI tools like LIME and SHAP to highlight feature importance. The study aims to discern user behavior patterns to improve ad targeting. Models are further optimized using parameter tuning, and ensemble strategies, such as soft and hard voting, are used to aggregate individual predictions, boosting detection accuracy. Performance metrics, ranging from accuracy to specificity, are used to assess model efficacy. Results show that ensemble methods, particularly soft voting, outperform other techniques in accuracy.

Keywords: User Behavior Detection, Online Advertising, Machine Learning, Parameter Tuning Ensemble Voting, XAI.

TABLE OF CONTENTS

	<u>Pages</u>
ABSTRACT	vi
TABLE OF CONTENTS	vii
LIST OF FIGURES.....	x
ABBREVIATIONS.....	xiii
1. INTRODUCTION	1
1.1 OVERVIEW	1
1.2 PROBLEM STATEMENT	2
1.3 RESEARCH MOTIVATION	3
1.4 RESEARCH OBJECTIVES	3
1.5 RESEARCH QUESTIONS	4
1.6 RESEARCH CONTRIBUTION.....	4
1.7 RESEARCH PLAN	5
2. LITERATURE REVIEW.....	6
2.1 RELATED WORKS.....	6
2.2 USER BEHAVIOUR.....	8
2.2.1 Types of Online Advertising.....	8
2.2.1.1 Search advertising	8
2.2.1.2 Social media advertising	9
2.2.1.3 Email advertising	11
2.2.1.4 E-commerce advertising.....	11
2.2.2 Types of User Behaviour in Advertising	12
2.2.2.1 Attention and engagement	12
2.2.2.2 Perception and attitude	13
2.2.2.3 Trust and credibility	14
2.2.2.4 Avoidance and ad-blocking.....	14

2.2.2.5 Click advertising	15
2.3 MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE.....	16
2.3.1 Algorithms in ML	16
2.3.1.1 Supervised learning (SL).....	17
2.3.1.2 Unsupervised learning (USL)	18
2.3.1.3 Semi supervised learning (SSL).....	18
2.3.1.4 Reinforcement learning (RL)	19
2.3.2 Machine Learning Pipeline	20
2.4 MACHINE LEARNING ALGORITHMS	21
2.4.1 Ensemble Learning (EL).....	21
2.4.2 Random Forest (RF)	22
2.4.3 Gradient Boosting (GB).....	23
2.4.4 Logistic Regression (LR).....	23
2.4.5 K-Nearest Neighbors (KNN)	24
2.4.6 Decision Tree (DT)	24
2.4.7 CatBoost.....	25
2.5 EXPLAINABLE ARTIFICIAL INTELLIGENCE	26
2.5.1 Explainability and Interpretability	26
2.5.2 Explainable Artificial Intelligence Methods	27
3. METHODOLOGY.....	29
3.1 PROPOSED METHODOLOGY DESIGN	29
3.1.1 Dataset Gathering.....	30
3.1.2 Data Analysis	30
3.1.2.1 Age distribution.....	30
3.1.2.2 Gender distribution.....	31
3.1.2.3 Ad effectiveness	31
3.1.2.4 Correlation analysis.....	32
3.1.3 Data Pre-Processing	33

3.1.3.1 Data quality verification.....	33
3.1.3.2 Data splitting	34
3.1.4 Proposed Classifiers.....	35
3.1.4.1 Random forest classifier.....	35
3.1.4.2 Gradient boosting classifier.....	36
3.1.4.3 Logistic regression classifier.....	37
3.1.4.4 K-nearest neighbors classifier	39
3.1.4.5 Decision tree classifier	40
3.1.4.6 CatBoost classifier	41
3.1.4.7 Ensemble model	42
4. RESULTS AND DISCUSSION.....	43
4.1 THE USED LIBRARIES.....	43
4.2 EVALUATION METRICS	44
4.3 MODELS PERFORMANCE WITHOUT PARAMETER TUNING	45
4.3.1 Evaluation of the Random Forest Model	45
4.3.2 Evaluation of the Logistic Regression Model.....	48
4.3.3 Evaluation of the Gradient Boosting Model	51
4.3.4 Evaluation of the KNN Model	53
4.3.5 Evaluation of the DT Model	56
4.3.6 Evaluation of the CatBoost Model.....	59
4.4 BASE MODELS COMPARISON RESULTS	61
4.5 ENSEMBLE MODELS RESULTS.....	62
4.5.1 Results using Soft Voting	62
4.5.2 Results using Hard Voting	63
4.5.3 Ensemble Models Comparison	65
4.6 DISCUSSION	65
5. CONCLUSION AND FUTURE WORK.....	66
References.....	67

LIST OF FIGURES

	<u>Pages</u>
Figure 2.1: The Most Widely Used Social Networks Worldwide as of January 2023, Ranked by the Number of Monthly Active Users.	10
Figure 2.2: ML Algorithmes Types.	17
Figure 2.3: Supervised Learning.	17
Figure 2.4: Unsupervised Learning.	18
Figure 2.5: Semi-Supervised Learning.	19
Figure 2.6: ML Pipeline.	20
Figure 2.7: Ensemble Learning [40].	21
Figure 2.8: Random Forest [42].	22
Figure 2.9: Logistic Regression [45].	24
Figure 2.10: Decision Tree [48].	25
Figure 3.1: The Proposed Methodology Design.	29
Figure 3.2: Users Age Distribution Histogram.	30
Figure 3.3: Users Gender Distribution.	31
Figure 3.4: Ad Effectiveness Label Distribution.	32
Figure 3.5: Dataset Correlation Matrix.	33
Figure 3.6: Random Forest Parameters Grid.	35

Figure 3.7: Gradient Boosting Parameters Grid.	37
Figure 3.8: Logistic Regression Parameters Grid.....	38
Figure 3.9: DT Parameters Grid.	41
Figure 3.10: Catboost Parameters Grid.	42
Figure 4.1: Random Forest Confusion Matrix.	46
Figure 4.2: Random Forest Classification Report.	46
Figure 4.3: Feature Importance Derived from LIME Analysis on a RF Model.....	47
Figure 4.4: Average Impact on Model Output Magnitudes for RF Using SHAP.	48
Figure 4.5: Logistic Regression Confusion Matrix.	49
Figure 4.6: Logistic Regression Classification Report.....	49
Figure 4.7: Feature Importance Derived from LIME Analysis on a LR Model.....	50
Figure 4.8: Average Impact on Model Output Magnitudes for LR using SHAP.....	50
Figure 4.9: Gradient Boosting Confusion Matrix.....	51
Figure 4.10: Gradient Boosting Classification Report.	52
Figure 4.11: Feature Importance Derived from LIME Analysis on a GB Model.	52
Figure 4.12: Average Impact on Model Output Magnitudes for GB Using SHAP.	53
Figure 4.13: KNN Confusion Matrix.	54
Figure 4.14: KNN Classification Report.	54

Figure 4.15: Feature Importance Derived from LIME Analysis on a KNN Model.	55
Figure 4.16: Average Impact on Model Output Magnitudes for KNN Using SHAP.	56
Figure 4.17: DT Confusion Matrix.....	57
Figure 4.18: DT Classification Report	57
Figure 4.19: Feature Importance Derived from LIME Analysis on a DT Model.	58
Figure 4.20: Average Impact on Model Output Magnitudes for DT Using SHAP.....	58
Figure 4.21: Catboost Confusion Matrix.....	59
Figure 4.22: Catboost Classification Report.	60
Figure 4.23: Feature Importance Derived from LIME Analysis on a CatBoost Model.....	60
Figure 4.24: Average Impact on Model Output Magnitudes for CatBoost Using SHAP. ..	61
Figure 4.25: Base Models Accuracy Comparison Before and After Parameter Tuning.	62
Figure 4.26: Soft Voting Confusion Matrix.	63
Figure 4.27: Soft Voting Classification Report.....	63
Figure 4.28: Hard Voting Confusion Matrix.....	64
Figure 4.29: Hard Voting Classification Report.....	64

ABBREVIATIONS

ML	:	Machine learning
AI	:	Artificial Intelligence
XAI	:	Explainable Artificial Intelligence
DT	:	Decision Tree
SL	:	Supervised learning
USL	:	Unsupervised learning
SSL	:	Semi Supervised Learning
RL	:	Reinforcement Learning
EL	:	Ensemble Learning
RF	:	Random Forest
GB	:	Gradient Boosting
LR	:	Logistic Regression
KNN	:	K-Nearest Neighbors
LIME	:	Local Interpretable Model-Agnostic Explanations
SHAP	:	SHapley Additive ExPlanations
SVM	:	Support Vector Machines
ACC	:	Accuracy
PRE	:	Precision
REC	:	Recall
F1-S	:	F1-Score
CR	:	Classification Report
CM	:	Confusion Matrix
P	:	Python

1. INTRODUCTION

1.1 OVERVIEW

According to online advertising definitions, a visitor on the internet means a person who visits any website or social media platform and takes part in its content in a particular way [1]. These two realms are therefore entirely new venues for human interaction where individuals, groups, and even governments can carry out social, political, and economic activities as well as information exchange [2]. In the era of big data and digital technologies, marketing and especially digital marketing are rapidly developing. This development is significantly supported by utilisation of various quantitative methods allowing for a lot of useful information from different marketing fields. Moreover, users are also able to peruse numerous pages, select links to read more about products, or spend time on the site. Businesses around the globe have been using these channels to attract consumers and establish productive marketing relationships with those clients. Therefore, one should observe how consumers behave and interact in a more careful manner. Only then can we obtain valid evidence about why they act as they do and why it is crucial for businesses to develop effective advertising campaigns targeted at the right people. This will encourage users to engage with their content. It enables companies to identify trends and patterns that can be used to refine their marketing strategy [3]. The creation of a digital advertising landscape in the modern era has changed the means by which businesses gain access to their target market. Nevertheless, the large number of commercials available online implies that customers could be turned off and choose not to engage with them. The amount of money companies invest in their advertising campaigns is a clear sign of how many people are interested in this sector; for example, according to 2016 data from Statista [4], about 524.58 billion USD was spent on advertisements. Social media ads are another form of advertising that is gaining huge traction. In fact, with an estimated investment of about 32.3 billion USD on desktop and mobile social media ads in 2016, as mentioned by Statista [5]. However, it brings up the question of whether this is an advisable choice for the company from a financial perspective. On top of that, marketers are always challenged with creating more effective and attractive ads on social media that can appeal to their potential customer base. In May 2021, Facebook was found to be the most visited site in the United States with a visitation rate of 71.8% among all sites people visit on social

media [6]. Facebook had taken over third place with 9.15% of sales. Instagram and Pinterest saw 3.82 and 12.4 percent market share, respectively [7]. The annual revenues of the Meta Family of Apps came to \$117.92 billion in 2021, more than the \$85.97 billion a year ago. As of the last ten years, revenues at the company soared to over \$114 billion [8].

There are two types of behaviors, the click one and the non-click one. From [9] The click-water behavior means clicking on a commercial. It is often the main indicator to gauge whether a marketing campaign has been successful. This is because higher click-through rates indicate that advertising content is suitable and good at attracting the audience's interest. Therefore, non-click behavior refers to the phenomenon of not clicking on ads which is also recognized in case any time they can favor another. Analysis of these data could provide insight into the needs of users, so data manipulations that could lead to more efficient use of advertisements. We can see from this analysis patterns in the behavior of users who don't click and their dislike of some types of ad content, or the more appropriate time of day for posting ads.

1.2 PROBLEM STATEMENT

The effectiveness of any advertising campaign is based on the quality of advertisements used. In this regard, advertisers should study how their target audience behaves and their preferences in the digital world. Unfortunately, methods that analyze user behavior, such as surveys or focus groups, are usually tedious and fail to yield positive outcomes for an advertiser. Selection bias is another problem where the results do not completely represent the target audience as a whole. Due to the development of ad-blocking technology, it becomes even more difficult to reach the intended audience. These ad blockers not only block ads, but they also prevent advertisers from gathering data about our likes and dislikes. One way of dealing with this problem is to utilize user behavior detection which can examine information from the likes of website analytics and social media footprints. In addition to creating impressive advertisements, user behavior detection can shield you from the ravages of click fraud and other forms of iniquity which might foil your advertising efforts completely. By preventing such practices, advertisers can be certain that their ads are reaching the desired audience and stretch their ad money to greater effect. The digital advertising world is changing rapidly, introducing new platforms and technologies so quickly that advertisers should find a way to keep abreast of these changes. With user

behavior detection, the system will be able to identify user trends and preferences that are emerging, enabling them to adopt appropriate strategies and remain relevant.

1.3 RESEARCH MOTIVATION

As technology continues to advance at an unprecedented rate, the world of advertising has undergone a significant transformation. One of the most prominent shifts in recent years has been the rise of advertising. With the widespread adoption of smartphones and the increasing amount of time people spend on their mobile devices, advertisers have recognized the immense potential in reaching their target audience. By user behavior detection has gained attention due to its advantageous nature towards advertisers and users. With personalized and targeted ads, advertisers can enhance the user experience, elevate the level of engagement, and eventually have more conversions that lead to increased ROI. Simultaneously, users would find it easier to engage with their own personalized advertisements, which will make the entire experience very satisfying for them. Moreover, the creation of an accurate yet efficient machine learning model for a behavior detection system also adds value to the evolution of the ML & AI field itself, as well as the digital advertising industry at large.

1.4 RESEARCH OBJECTIVES

The purpose of the advertisement user behavior detection study was to build a precise and efficient machine learning model for processing user behavior data and preferences at lightning speed. This makes it possible for ads to be custom-tailored to an individual's behavior pattern and likes. By calculating the user's behavior, the system needs to process large numbers of data and manage complex algorithms beginning with instant assessment and ad optimization. For ad campaigns to be transparent the model is based on explainable artificial intelligence (XAI) that people can understand. It is a process that enhances the entire user experience by providing more appropriate and stimulating ads, as well as giving advertisers a better ROI on the conversions that result from their targeted advertising. The goal is really more about advancing in the domains of ML and AI, which can be achieved by creating precise, efficient, and explainable ML models for user behavior detection.

1.5 RESEARCH QUESTIONS

- a. How can user behavior detection be used to improve the user experience and increase the return on investment for advertisers?
- b. How can ML be used to accurately and efficiently analyze user behavior and preferences in relation to digital advertising?
- c. How can an ML model be developed that is capable of handling large amounts of data and complex algorithms, providing real-time analysis and optimization of ads based on user behavior?
- d. What impact do personalized and targeted ads have on user engagement and conversion rates, and how can this impact be measured?
- e. What are the most effective methods for measuring the effectiveness of personalized and targeted ads created using user behavior detection techniques?

1.6 RESEARCH CONTRIBUTION

The contribution of this research is:

- a. The formulation of an ML-based methodology to scrutinize user behavior and preferences concerning digital advertising, complemented by the optimization of advertising initiatives via EL methodologies.
- b. Through the employment of ML algorithms such as RF, LR, and GB, and the integration of EL techniques like VotingClassifier, this research endeavors to craft ads that are both impactful and align with the inclinations of the target demographic, fostering higher conversion rates.
- c. This investigation extends the frontiers of knowledge in the realm of tailored and pinpointed advertising by identifying nascent trends and inclinations in user behavior and appraising the efficacy of such tailored advertising strategies.
- d. A significant highlight of this research is the integration of XAI to ensure that the ML models are not only effective but also transparent, allowing stakeholders to comprehend and trust the underlying decision-making processes of the models.
- e. The ultimate aspiration is to elevate the user experience and augment the return on investment for marketers by orchestrating advertising campaigns that are more precise, agile, and resonate with their intended audience.

1.7 RESEARCH PLAN

The research blueprint for this thesis unfolds in a structured manner.

Literature Review: the study will engage with prior scholarly work on user behavior in online advertising, exploring various ad modalities. It will also dissect various user behaviors and their impact on advertising while diving deep into the domain of click advertising. The role of ML and AI in enhancing ad strategies will be thoroughly examined, focusing on specific algorithms and ensemble techniques.

Methodology: the research design will be unveiled, detailing the dataset sourcing, its analytical framework, preprocessing, and the classifiers being adopted, with an emphasis on ensemble methods.

Results and Discussion: will introduce the evaluative framework, tools, and libraries, followed by performance assessments of models, both pre and post-optimization. The section will culminate in a comparative analysis of ensemble techniques.

Conclusion and Future Work: encapsulating the core findings, their implications, and prospective research avenues in the domain.

2. LITERATURE REVIEW

2.1 RELATED WORKS

A review of the literature on user behavior in ML refers to an assessment of the previous studies conducted on this subject. The review aims to examine the utilization of ML algorithms in modeling and comprehending user behavior, and will analyze the principal discoveries and patterns identified in this field of study. This study [10] looks into how big data technologies may be used to forecast customer behavior on social media sites. The study use mathematical modeling created by ML to anticipate customer behavior by examining consumer activity on social media using numerous characteristics and criteria. The decision tree (DT) model is the one that predicts consumer behavior on social media sites most accurately, according to the results. According to the survey, customer deviation varies among social media platforms from 12.22% to 99.51%. The largest root mean square error is 156556.45293 and the lowest is 20691.7870. The model's ACC varies between 0.02238 and 0.98292. The purpose of [11] paper is to exhibit possibilities of logistic regression identify the utilization in prediction user in order to detect the click fraud in online environment characterized by particular demographic features. The only variables exploited in the model by the mean of stepwise regression are variables with significant influence. The impact of particular factors is quantified via odds rate that are used for the identification of areas of interests typical for women, men and for considered age categories. Prediction quality of models is value by the set of classification measures arising from confusion matrix that is generally acceptable in machine learning. The study demonstrates the usefulness of logistic models for estimating the probabilities of an internet user belonging to a target group. A technique to eliminate click fraud from ad networks was introduced in [12], where every click can be judged as suspicious by using a comparison with multiple sets of standards, which could prevent the spread of the problem. The tested agents performed well when placed into a testing environment, and they identified different types of attacks without a single doubt. In practice, the authors recommended re-weighting each criterion because if wrong weights are chosen, then not only will this affect the improper detection of attacks, but also false alarm rates and type 1 errors would increase exponentially. In addition to the previous point about cyber fraud, this method serves as a possible solution for the huge problems facing internet advertisers

on how to reduce their vulnerability due to high percentages lost from such unauthorized activities. To develop a comprehensive contextual advertising strategy, we examined ways to refine company page targeting based on consumer behavior on Facebook, specifically focusing on age and gender. An algorithm proposed in this research paper [13] uses decision trees to guide and facilitate the regulation and optimization of marketing strategies for a business enterprise by applying statistical training principles to commercial objectives. The algorithm aims to maximize the revenue generated by campaigns, achieved by simulating engagement with a company page on Facebook. The developed method was tested with data from a Facebook advertising case, and simulation results show that the algorithm identifies target groups that are highly profitable and outperform industry benchmarks. The research confirms the efficiency of the proposed approach in establishing decision trees for customer engagement while keeping company priorities in focus. The approach suggested in this study employs association rule mining and customer feedback categorization to suggest projects to freelancers. To categorize completed works as good or negative, they begin by compiling the work histories of freelancers and utilizing LR and Linear SVM models to assess the mood of customer comments. Finally, we utilize association rule mining to determine which skill sets freelancers frequently employ across both types of successfully completed assignments. We then employ set operations to find employment that correspond to the positive frequent skillsets while eliminating those that do not. Lastly, to produce a more precise suggestion, we use a collaborative filtering algorithm that takes into account client ratings, minimum budget/hourly rate, deadline, and re-hire. Our tests on actual datasets from several online marketplaces show that the ACC of our suggested method's recommendations for acceptable jobs is 83.40% (LR) and 84.03% (Linear SVM). In the research presented in this paper [9], the Search-based Interest Model (SIM) is introduced, employing two distinct search units: the General Search Unit (GSU) and the Exact Search Unit (ESU). These units collaboratively work to derive user interests from extensive sequential behavior data. The GSU initially performs a broad search using query information from a candidate item to yield a relevant Sub User Behavior Sequence (SBS). Subsequently, the ESU precisely models the relationship between the SBS and the candidate item, enhancing SIM's ability to accurately and efficiently capture lifetime sequential behavior data. This novel approach has been effectively integrated into Alibaba's display advertising system since 2019, resulting in notable improvements of

7.1% in click-through rate (CTR) and 4.4% in revenue per mille (RPM). Currently, SIM manages the bulk of traffic in Alibaba's operational system. This essay focuses on the problem of click fraud in internet advertising, which is dangerous for e-commerce and the advertising sector [14]. Although there have been improvements made to traffic filtering methods, efficient fraud detection algorithms are still required to safeguard online advertising firms. In order to discover IP, we have studied click patterns over a dataset that deal 200 million clicks over four days. The goal was to assess the trip of a user's click across their portfolio and flag IP addresses that produce lots of clicks. The ACC was 98% because to the implementation of the LightGBM ML algorithm, a technique similar to GB and DT. The research emphasizes the significance of creating efficient fraud detection algorithms for online advertising organizations and relies on a literature analysis to verify the results.

2.2 USER BEHAVIOUR

The modern consumer economy places a particularly great onus on advertising. Marketers, as well as advertisers, need to appreciate the connection between user behavior and advertising better than ever before. After seeing an ad, people's activities, emotions, and reactions all come under the heading of user behavior. This includes components including attention, perception, trust, and response. Let's take a look what kinds of user behavior in advertisements there will be. Whether people's attitude to ads or their involvement with them. We are to talk about how culture and demographics affect user behavior, and what new trends in this area may imply for those who sell things despite all opposition. Experts in marketing can devise even better strategies for their target audience when they understand the psychology behind a person being advertised at. And because this same knowledge may be employed by policymakers to regulate what advertising companies do and how ads are delivered--not just so their customers won't see dodgy or dangerous ads, but also offer some protection to them [15].

2.2.1 Types of Online Advertising

2.2.1.1 Search advertising

Also known as search engine marketing (SEM), is a form of online advertising that involves displaying text ads within search engine results pages, such as Google Ads [16]. The ad is shown according to what users type in, and then, these pages are arranged

together either up or down. Normally these ads are marked “sponsored” or “ad” signs of paid listings and thereby set off from organic search results. This kind of advertisement is especially valuable in that it is targeted; businesses can reach consumers who are already interested in their products without all the "maybe" that go along with newspapers or billboards. Use of search engines such as Google AdWords allows a business to exert control over where and when the ad is displayed based on the keywords it chooses, as well as enabling targeting by geographic location or device among other criteria. One effective method this advertisement can be used is driving traffic to their site and generating leads for sales; however, it may prove fiercely competitive, and the bidding process should be closely monitored so that businesses are making the best use of their advertising investments. What's more, this type of online advertisement demands constant optimization actions for your advertisements not to become stale too soon but to keep delivering expected results in a proper way, along with being flexible enough so that if there is anything wrong, you would catch that early enough to fix it. Another essential part of advertising online is Search Engine Optimization (SEO), where, to rank higher in search engine results pages (SERPs), website content and structure are optimized organically [17]. The main purpose of SEO is to make sure that websites can be seen by people looking for information using keywords in search engines and to attract more traffic naturally without paying for it. There are various methods in SEO including meta tags and descriptions optimization, increasing the speed of a website as well as its mobile-friendliness, creating high-quality content, and developing backlinks [18]. These activities enable businesses to increase their visibility on the web, thereby bringing more traffic without requiring them to pay any amount for adverts. SEO does not deliver instant results and calls for an ongoing effort, making it a long-term strategy. By complementing SEO with search advertising, businesses can have a two-tiered approach to internet search engine marketing, through organic rankings as well as paid advertising [18].

2.2.1.2 Social media advertising

When it comes to social media advertising, sponsored posts and display ads are offer a prime advertising space for businesses. By placing ads within popular mobile apps, you can reach a large and engaged audience. For example, you can run ads on social media platforms like Facebook, Instagram, or Twitter. Such targeted advertising ensures that companies connect only with the subset of potential consumers who are most likely to be

interested in what they sell or provide based on their demographics, interests, and behaviors. There are various types of advertisements that can be featured on social media: display ads, video ads, carousel ads, sponsored posts, and more. These adverts appear on users' newsfeeds or other areas within the website such as stories or sidebars [19,20]. Referred to as "sponsored" or "ad" they are usually differentiated from organic content [21]. One benefit of this form of marketing is the ability to engage with audiences through interactive content like surveys, quizzes, and competitions [22]. By using this strategy, companies can enhance their sales, generate leads, and build brand awareness [23]. online advertising allows businesses to track the performance of their ads in real-time and make adjustments as needed. This makes it possible for startups to fine-tune their campaigns and quickly identify which strategies are working and which ones are not [20]. Even though it offers several advantages, social media advertising does present challenges, Content companies to create content that resonates with your target audience, addresses their pain points, and provides solutions to their problems. By doing so, you can establish yourself as a thought leader in your industry and create a loyal following of customers, by creating materials that attract your target audience, address pain points, and provide solutions to problems. Doing so can help you establish yourself as an authority figure in the field.

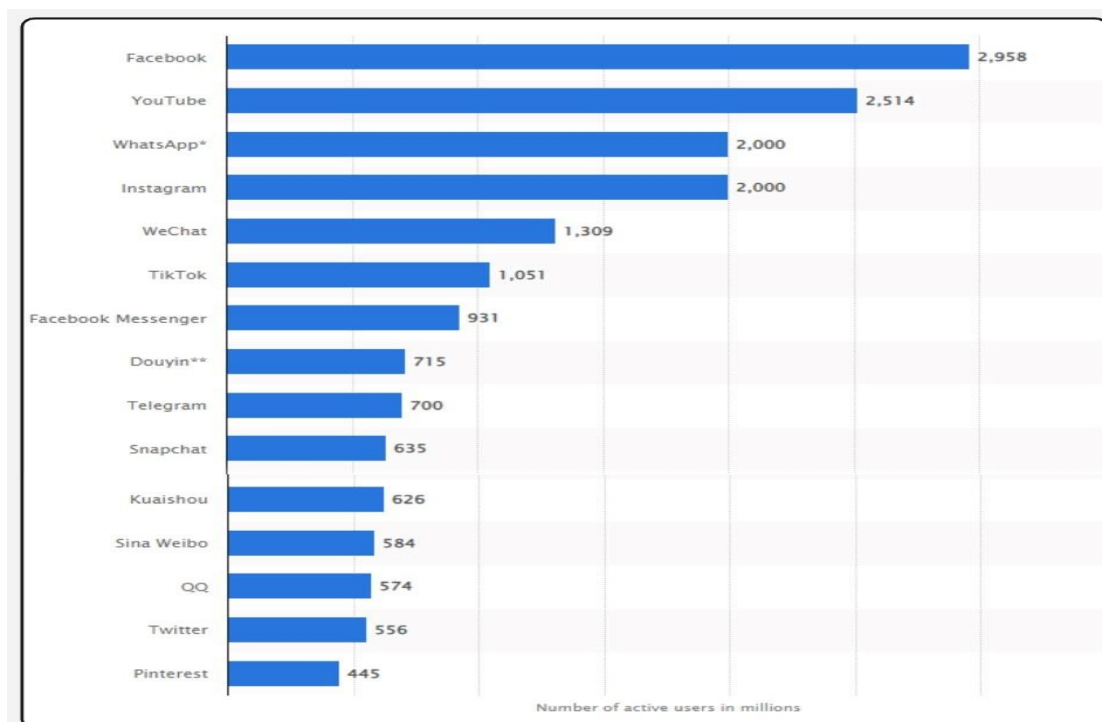


Figure 2.1: The Most Widely Used Social Networks Worldwide as of January 2023, Ranked by the Number of Monthly Active Users.

2.2.1.3 Email advertising

Electronic mail promotion is a form of web advertising that sends out sales letters to a list of subscribers based on targeted advertisements for the promotion of a product or service [20]. email advertising can be an effective way to generate leads and sales, but it can also be easy to spam people if you're not careful. When choosing which type of online advertising to invest in. Generally, it is employed by corporations as part of building connections with subscribers and delivering leads and sales to a business process. Personalization is another advantage of email marketing that comes with a high level of customization. Segmenting the email list according to demographic information, interests, and behavior will enable the business to customize its messages based on audiences. In addition, email marketing is considered to be a low-cost method due to its cost per impression and the fact that it can reach a wide audience with just one email [24]. On the other hand, there are also downsides to email marketing, one of which is the fact that subscribers could possibly mark the emails as spam, resulting in a negative impression on the sender's reputation and deliverability [25]. In order to avoid these pitfalls, companies should ensure that their campaigns abide by anti-spam legislation and provide quality content that adds value to their readers. Additionally, they need to keep track of how well each individual campaign performs so they can tweak elements accordingly - making sure there is no waste on investing in underperforming campaigns while making the most use out of them for successful ones.

2.2.1.4 E-commerce advertising

Electronic commerce, also known as e-commerce, is a subgenre of online advertising designed to draw prospective clients toward individual e-commerce stores or general online retail sites like Amazon or eBay. The primary aims of e-commerce advertising are to increase online sales and the volume of traffic that visits the websites where these products and services are being sold. According to [10], many forms of e-commerce advertising such as search, display, and social media adverts have been identified. For instance, displaying targeted ads to users searching for products or services directly yields highly qualified traffic to a website, thus improving conversions. One of the key parts is managing and optimizing PPC campaigns to generate earnings for enterprises as they operate. Display targeted ads to users who are actively searching for products or services, PPC can drive qualified traffic to a website and increase conversions. However, it's

important to carefully manage and optimize PPC campaigns to ensure that they are delivering a positive return on investment. By following best practices, avoiding common mistakes, and leveraging advanced strategies and tools, businesses can achieve even better results. The types of social media ads, such as sponsored posts or display adverts on sites like Facebook and Instagram, are called social media ads. The major advantage of e-commerce advertising is that it is capable of achieving a very specific audience. Different features such as interests or behaviors among different demographics can be used to judge the user's attitude towards a product. Moreover, e-commerce advertising is a great low-cost option because advertisers only pay when the ads are clicked or shown. However, e-commerce advertising must carry some disadvantages. But since it can employ a big base of potential online customers who shop everywhere, the competition among online shoppers could be very intense. Advertisers must ensure that ads are original and provide real value for potential customers. These efforts must be monitored and optimized to ensure they really do work to achieve their intended results and produce the desired ROI.

2.2.2 Types of User Behaviour in Advertising

In the field of advertising, user behavior can manifest itself in various ways, and it is important to appreciate these diverse manifestations if one wants to make successful advertising campaigns. The crucial user behavior that has impacts on advertising consists of attention, perception, attitude, response, trust, avoidance, as well as cross-cultural differences, demographic differences, emotion, and cognitive processing.

2.2.2.1 Attention and engagement

User attention and engagement are central ingredients of users' behaviors in advertising. Attention is the users' focus on an advertisement influenced by a variety of elements including the format, placement, and the content [26]. One critical factor that has a substantial impact on attention is the length of the ad. longer ads are seen as less interesting, whereas shorter ones tend to grab more attention because it is likely that they will be watched entirely. There have been studies that found out that pre-roll video ads, which are 15 seconds or less, have better viewership than those that are longer. Also affecting attention, and this is still based on where the ad appears, is the placement of the advertisement. When advertisements are placed in areas users are likely to focus, such as at the top or center of a web page, they get more attention compared to less prominent

placements. Moreover, the relevance of the ad to the user plays a critical role in drawing their attention. Users will be more inclined towards ads that are pertinent to their interests or needs. The second indicator of ad effectiveness is engagement, which entails the user's voluntary interaction with an ad. Ways to gauge engagement include observing the duration for which users interact with an ad, the count of clicks or shares, and also the emotional response from it [27]. The factors that influence engagement are the format of the ad, its message, and tone. Interactive ads, like quizzes or games, tend to be more engaging because they compel users to interact with the ad. Advertisements that use humor, emotional appeals, or storytelling approaches are more likely to create a strong emotional bond between the user and the ad [27]. The message and tone of the ad are also key factors in determining engagement. Ads that articulate a clear and precise message while using a friendly tone are more engaging.

2.2.2.2 Perception and attitude

Relevance, entertainment value, authenticity, and transparency are among the numerous factors that can determine user attitude toward ads. For instance, users are likely to have a positive attitude towards ads when they are interesting and provide value benefits. A commercial seen during a favorite television program or on a well-respected website can get more favorable reactions than one viewed in an undesired or unfamiliar context. Also, user-generated content (UGC) such as social media posts and product reviews shapes people's attitudes toward advertising. If the user-generated content (UGC) related to the ad's message is negative, it is much easier for people to accept a new product or idea and want to buy it. Once ad fatigue begins to set in and users start feeling overwhelmed or irritated by the sheer number of ads, they tend to have an unfavorable impression of advertising. Over time, as messages come at users from all sides whether they're interested or not, or even when the ads are irrelevant, people may lose interest in them. After all, exactly how ads target and personalize themselves matters when it comes to users' attitudes towards ads. Most consumers may find personalized advertising acceptable as long as it meets their needs in some way, and therefore prefer it over non-targeted or uninteresting ads. Personalization does cross a fine line into privacy, but this boundary is violated too often by companies. Advertisers must be careful not to cross the line at any time.

2.2.2.3 Trust and credibility

Transparency in advertising is an important way of building credibility. Users are most likely to trust advertisements that are clear in their messaging and in their goals. Also, if they can give accurate product or service information, they can be sure to trust them. Another advertising format that influences people's attitude toward advertising is the credibility level. Those which may be integrated into a site or application from an installation of ads in the web page above content, are deemed more reliable than banners that appear separately from content. These include native ads which act as part of a site or application's own content [29]. One of the factors that also affects users' opinions about credibility is the ad placement. Advertising has gained popularity in recent years, and adtech startups are leveraging this trend to provide innovative solutions for brands and publishers. Native ads blend seamlessly with the content on a website or app, providing a non-disruptive and engaging experience for users. One other contributing factor, which is social proof, is that an individual is affected by what others believe and do, can help shape the consumers' perceived ad credibility. Thus, it is often the case that ads containing social proof items (like customer reviews or testimonials) are perceived as trustworthy and reliable. The last thing to consider is how credible users perceive the brand whose ad they are looking at to be. On the contrary, unknown or untrustworthy brands' ads can be treated with doubt and suspicion.

2.2.2.4 Avoidance and ad-blocking

Having come up alongside the expansion of digital advertising, the act of ad-blocking has also become popular. To this day, many users prefer to use ad-blocking software in order to keep away from irrelevant and intrusive advertisements [30]. This development constitutes a major difficulty for advertisers as they need to make efforts to avoid being regarded as intrusive or disagreeable [31]. Advertisers need to create non-disruptive and engaging ad experiences to combat ad-blocking and ensure their messages reach the intended audience. We will also find out what this means for ad revenue. And there are also some ideas in our bag for how to not only solve the problem, but make your own product capable of projecting a friendlier attitude as well [32]. In the previous few years, most users put ad-blocking apps on their phones which was to not only see any ads at all. When advertisements stray from an individual's area of interest or requirements, they can be perceived as intrusive or annoying. While advertiser offers numerous advantages, it also

faces challenges such as ad-blocking. Many users install ad-blocking software on their devices to avoid intrusive or irrelevant ads. Advertisers need to create non-disruptive and engaging ad experiences to combat ad-blocking and ensure their messages reach the intended audience. Ad quality also plays a role in users' ad-blocking behavior. Ads that are of low quality, intrusive, or disturbing are prone to be blocked by users [33]. But ads that are creative, not invasive but engaging are less likely to encounter such dire consequences. What ad-blockers might ignore is that they can also block a substantial portion of the advertising income as you can't give them any ideas or clicks if the ads are not visible. For your business or blog if one of its most important sources is advertising, then it might be time for a fresh approach--taking into account people's behavior and responding accordingly. You have to create more interactive content in this area without being too intrusive. Or is it better to explore some other forms like selling products, because stopping ad-blocking could provoke frustration on both sides from our users and advertisers alike.

2.2.2.5 Click advertising

Click advertising, commonly known as pay-per-click advertising (PPC), is a widely used form of Internet marketing designed to cater for whatever specific set of user needs and objectives can be identified. In this mode, an advertiser pays for a user clicking the advertisement placed on his or her behalf. For general search engines, it can also be found in various places including social networks, e-commerce sites. Click ads target specific keywords, geolocations, and even demographics. As a result, they are the epitome of accurate marketing today. Advertisers use click ads to direct traffic, increase engagement, and, in the long run, raise conversion rates.

- a. Pay-per-click (PPC) Ads: These ads involve you paying per click, with a user clicking on them. Often, PPC ads are shown on search engine results pages and social media platforms [34].
- b. Display ads are clickable ads that show up on websites and applications in the form of images or videos. Based on users' browsing preferences or demographic details, display ads could be customized for a targeted audience.
- c. Native Ads: These are clickable ads that are intended to look like the rest of the content. They tend to be found on social media platforms and among other sponsored content on websites [34].

- d. Video Ads: Some video ads are interactive and are often displayed before or during online video content, for instance on YouTube videos or streaming services like Hulu.
- e. Shopping Ads: People using shopping ads can directly purchase the item by clicking, for example, on Amazon or Google Shopping sites.
- f. When it comes to app install ads, they are typically clickable and attractive in order to influence the user to download and install a mobile application. They may be found on different social media platforms, search engines, as well as in apps.
- g. Moreover, rich media ads are more advanced than simple clickable advertisements and include click-throughs that use unique features like videos or animations to encourage user interaction.

2.3 MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

The idea of artificial intelligence, according to Demis Hassabis, the CEO of DeepMind, a Google startup of AI, is defined as the field of creating machines that are intelligent. This definition fits well because AI is a general name for many applications, including categories like ML and deep learning, which use AI practically in reality. The use of computer systems to process a lot of data, learn, and improve with programming using ML makes it the most widely used type among all other AI types and also forms many bases of some AIs which marketers utilize. In general, a machine learning system is trained by applying a set of training data to establish a correspondence between input and output. The more data points the system deals with, the better it becomes with practice.

2.3.1 Algorithms in ML

To study ML and look at customer behavior, data is vital. We will delve into the importance of algorithms and their role in the ML pipeline in this section. Machine Learning (ML) algorithms refer to the way an AI system functions, usually producing output values given input data [35]. ML algorithms are classified differently, but there is always a categorization based on purpose with some subsets for each type. I had chosen the typical approach of categorizing algorithms into four main types – non-modifiers, simple modifiers, complex modifiers, and mixed compound types.

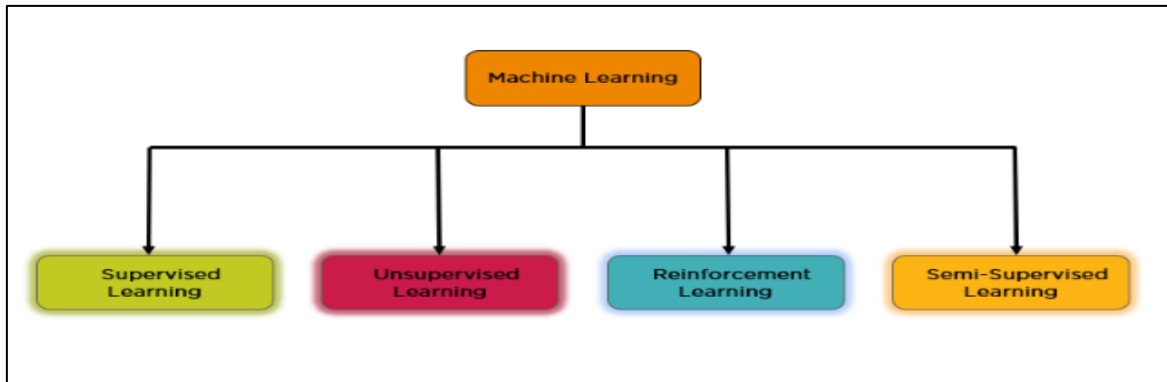


Figure 2.2: ML Algorithms Types.

2.3.1.1 Supervised learning (SL)

Supervised machine learning is a process through the training data of mapping inputs to specific output, estimating unknowns based on labeling samples. The objective of supervised learning technique is to build a model with distinguished features and predefining labels with a known class, In which desired solutions are available and often referred to as labels [36]. then using this model to classify or predict a new data with unknown class. The most well-known SL for classification is image recognition. For instance, an algorithm may be trained to identify images with cars and motorcycles through numerous example images indicating whether they represent a car or motorcycle. For instance, there is another SL task referred to as regression, which involves predicting a continuous numerical output [36]. Predicting rainfall based on features such as temperature, humidity, and wind speed in a given region is a classic example of regression. Consequently, based on these patterns and relationships between the features and the amount of rainfall learned, an application can use the extracted results to predict similar aspects for some other region that have not been seen previously.

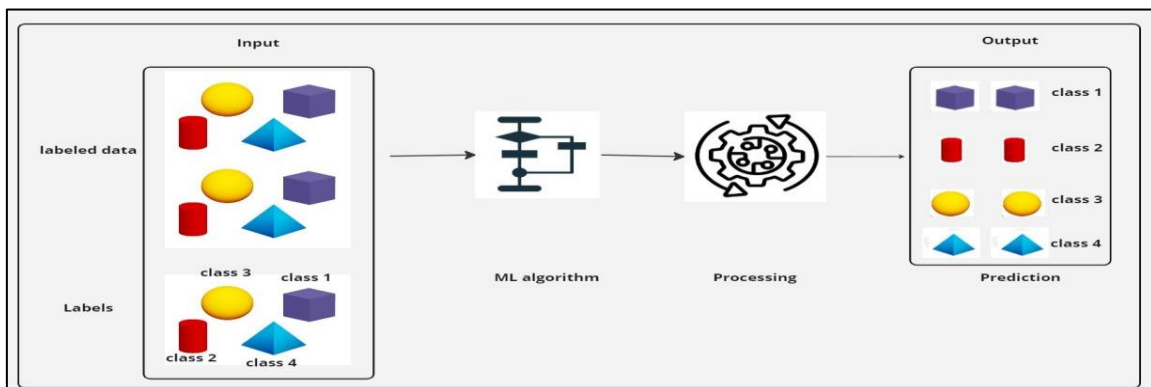


Figure 2.3: Supervised Learning.

2.3.1.2 Unsupervised learning (USL)

In USL (Fig.4), the input data does not have any predefined labels [36]. This is also referred to as unlabeled data, and unlike in SL, the algorithm tries to learn without a teacher [36]. An example of USL is anomaly detection. Imagine a dataset of credit card transactions, some of which may be fraudulent. An anomaly detection algorithm can analyze the patterns in the data and identify the transactions that deviate significantly from the norm, without any prior knowledge of what a fraudulent transaction looks like. This can help financial institutions detect and prevent fraudulent activities.

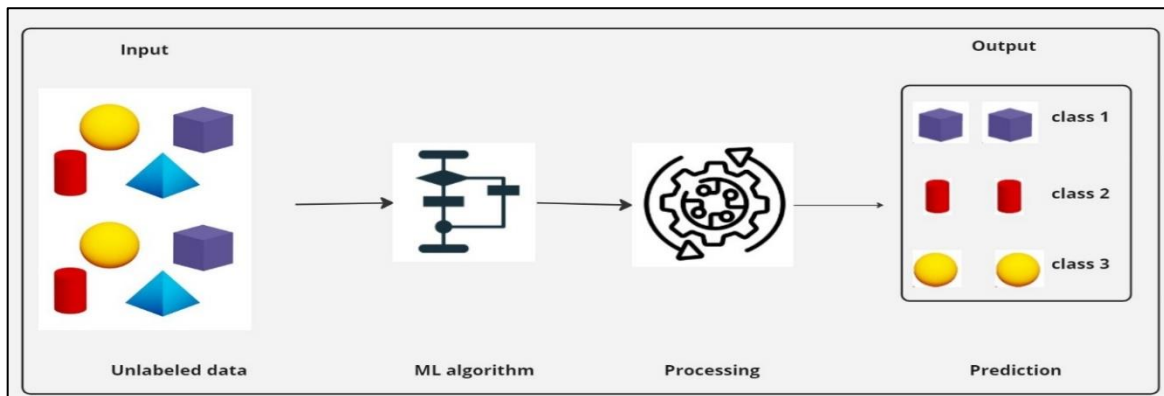


Figure 2.4: Unsupervised Learning.

2.3.1.3 Semi supervised learning (SSL)

SL and USL are two common types of ML algorithms that rely on labeled and unlabeled data, respectively. However, there is another class of methods named semi-supervised learning (SSL) that functions by combining labeled and unlabeled data, as given in reference [36]. An example where SSL is applied is natural language processing (NLP), and it means a model can be trained on a small collection of labeled data to a significantly large set of unlabeled data. In this case, once the model has learned the basic language rules such as grammar and syntax from a small dataset containing only labeled information, it applies those rules to predict what else might exist within much larger unsupervised datasets.

Semi-supervised learning is also a good method to use in anomaly detection systems because it helps differentiate normal and unusual events. Thus, the situation may be that most of the data remains unlabeled, but a smaller part of it is labeled as normal or abnormal [36]. Then, using labeled data, the model learns to identify what is considered normal behavior and then applies this knowledge when trying to find anomalies among the

unlabeled data [36]. SSL is also shown in action by Google's reCAPTCHA system, where human verification is required through object identification in images. From this identified data, the system improves its ability to detect objects within images automatically, which is one important part of stopping automated spam and bots.

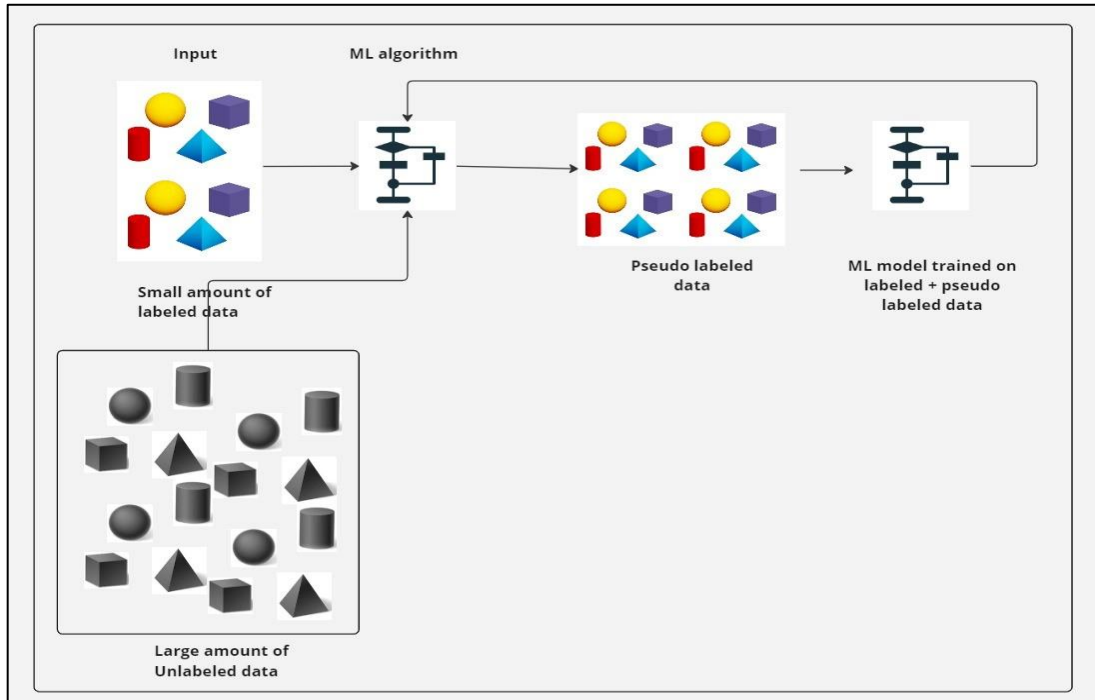


Figure 2.5: Semi-Supervised Learning.

2.3.1.4 Reinforcement learning (RL)

In order to learn more effectively, machine learning uses data from other sources, such as artificial intelligence. RL, a main field in machine learning, stands out from all others because of the great differences in how it comes by knowledge. This is how the RL algorithm learns—through interaction with the environment, getting rewards or punishments. One application of RL, for example, is to design autonomous driving systems that will guide a course through obstructing obstacles as well as patterns of traffic on roads. For example, applying deep learning methods to control cars with nothing more than video inputs through the cameras on the roofs of cars. Once they had gotten the hang of learning, these models could guide themselves over all kinds of new terrain seldom seen before understanding such detailed information as pedestrians leaving sidewalk curbs for various levels of traffic flow; they knew what conditions of speed were needed to slow down, whether to stay in the middle of the road or pull off on to a shoulder so other cars could pass them, etc. Prompts for users to respond quickly when conditions change, such as rapid

acceleration, skidding, stop-starts, and abrupt course changes can only get you through complex intersections safely if there are changes in multiple lanes. And blind curves, signs in different languages, traffic lights of various designs, and surfaces are examples of conditions where the intersection might have to be navigated by a self-driving car all without human input whatsoever! Another example: Recommender systems which work on applications such as Netflix or YouTube that understand user preferences through an analysis of viewing habits, thereby offering content similar to what they might enjoy.

2.3.2 Machine Learning Pipeline

Built from multiple interconnected fully automated methods, ML pipelines work on data operations of all kinds. To meet the needs of large-scale processing of data, these pipelines are typically found in machine learning systems [36]. The pieces that make up the pipeline do their work without getting in the way of the other components. Transferring the processed data to the other parts or saving it to the database for further analysis. It is a great opportunity for maintenance and development since when a part breaks, the system as a whole is not affected. Thus, the other components keep working. Typically, the ML pipeline consists of defining the problem, collecting and understanding the data, then comes pre-processing and model training, and finally, result visualization for interpretation. This pipeline is generic but could be tailored as per specific ML projects and can combine various algorithms or techniques to get the targeted results.

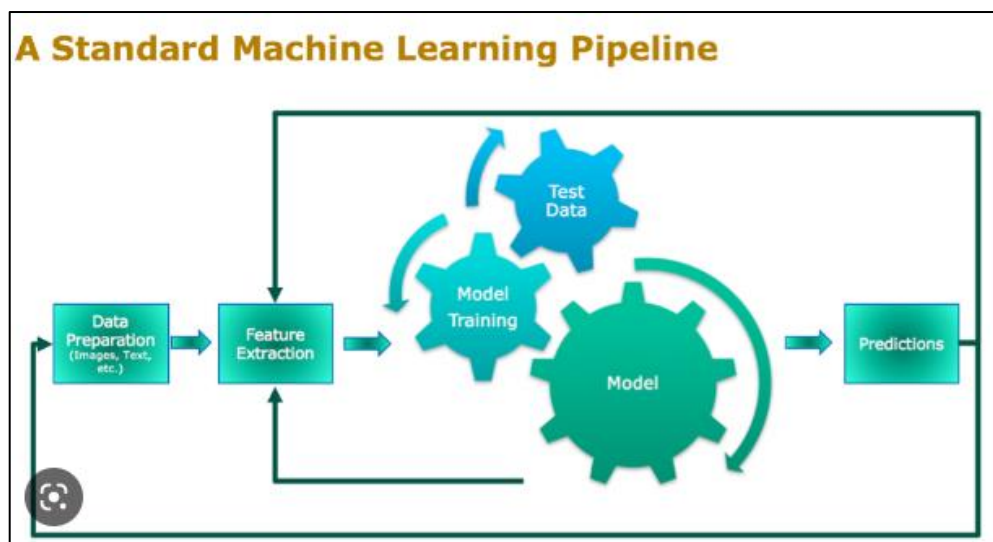


Figure 2.6: ML Pipeline.

2.4 MACHINE LEARNING ALGORITHMS

2.4.1 Ensemble Learning (EL)

An ML approach known as Ensemble Learning blends different models to elevate the accuracy of predictions. The underlying principle behind EL is that integrating predictions from multiple weak models can result in a stronger model that performs better on unseen data [36]. This has gained popularity as it improves accuracy and reliability of predictions while handling intricate data and non-linear relationships between input variables and target variables. There are various types of EL methods such as bagging, boosting, voting, and stacking, among others.

- a. One such approach is Bagging: short for Bootstrap Aggregating [37], which is characterized by training multiple models independently on different subsets of the training data and then combining their predictions.
- b. Another method is Boosting: the process iteratively introduces new models into the ensemble to correct the errors made by the preceding ones [38].
- c. This is known as voting: the combining of many models' outputs to produce a single prediction. Hard and soft voting are both examples of voting techniques. An individual vote, which is an outcome of a hard vote, will be selected as the final one based on the simple majority decision. However, the mean probability output of all individual models will be used for a soft vote to determine a final forecast.
- d. A method called Stacking: or Stacked Generalization [39], has been developed to train a collection of models and then use the output of these predictions as inputs to another meta-model, which then outputs the final prediction.

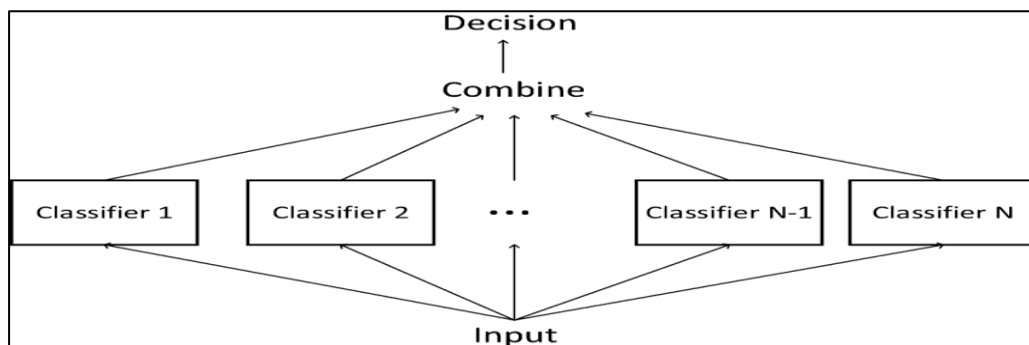


Figure 2.7: Ensemble Learning [40].

2.4.2 Random Forest (RF)

The RF algorithm is a meta-algorithm that combines multiple decision trees to form a powerful and robust classifier. The basic idea behind Random Forest is to build a collection of decision trees on randomly selected subsets of the training data, and then combine their predictions to obtain the final classification [41]. At each iteration, Random Forest randomly selects a subset of the features and a subset of the training features. The probabilities at each tree's leaves are averaged to determine the likelihood that an input belongs to a specific class. Unlike traditional DTs, each tree in the forest is grown independently and randomly. The data set used to train each tree is a randomly selected subset of the original training data, with replacement. At each node, a random subset of attributes is used to determine the best split, rather than considering all attributes. One of the strengths of Random Forest is its ability to handle complex decision boundaries and nonlinear relationships between the features and the target variable. It can also handle missing data, outliers, and irrelevant features, by randomly selecting subsets of the features. Instead, the RF algorithm is preferred as it retains the strengths of DTs while overcoming some of their limitations. RF reduces the risk of overfitting as the output of the classifier depends on the entire set of trees rather than a single tree. Randomness is introduced in the creation of each tree, preventing the classifier from memorizing all examples in the training set. Regularization techniques can also be applied to the trees in the forest, further reducing the risk of overfitting. However, like DTs, RF has a bias towards variables with many categories.

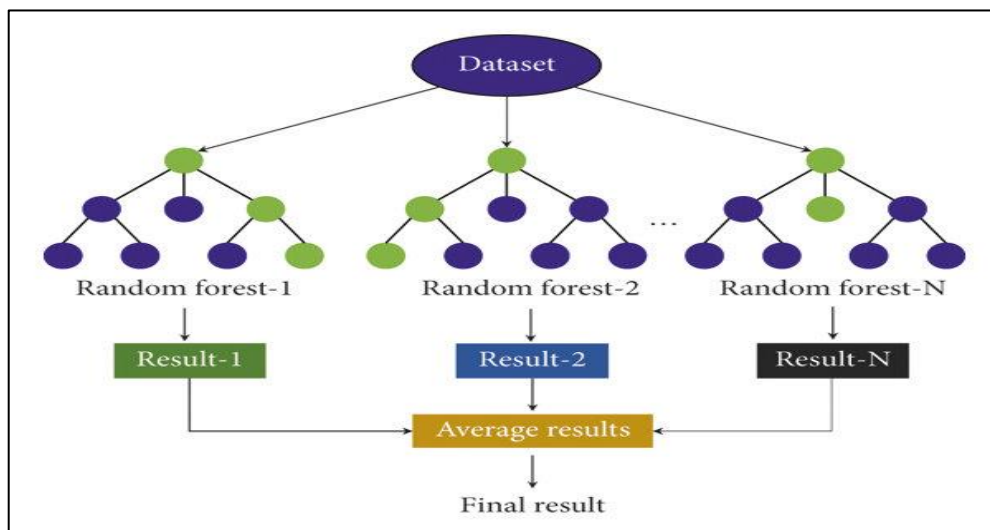


Figure 2.8: Random Forest [42].

2.4.3 Gradient Boosting (GB)

A ML approach called GB is used to create prediction models by fusing a number of weak models [43]. This specific boosting technique functions by continuously adding new models to the ensemble while fixing the shortcomings in the models that came before it. GB's fundamental premise is to train a series of weak models typically DTs on various subsets of the training data. The original data is used to train the first model, while the difference between the predicted and the actual values has been shown, the target variable are used to train subsequent models. GB calculates the gradient of the loss function for each iteration in relation to the predictions of the current model. Then, using this gradient, the predictions are updated in a way that minimizes the loss function. GB can build a robust ensemble model that performs well on unobserved data by carrying out this process iteratively. It has the capacity to handle complicated data and non-linear correlations between the input and the target variables, which is one of its benefits. As it may stop adding models when it starts to perceive declining results, it is also less prone to overfitting than other ML methods.

2.4.4 Logistic Regression (LR)

A supervised learning algorithm that models the relationship between the independent variables and the probability of belonging to a particular class, which converts a linear combination of input variables into a probability value between 0 and 1, is used to forecast the likelihood of an event occurring based on a collection of input data [44]. Finding the ideal model parameter values that minimize the discrepancy between the projected probabilities and the actual labels of the training data, In this type of learning, the algorithm is trained on a labeled dataset. The labeled dataset consists of input features and their corresponding output labels. The algorithm learns to map the input features to the output labels and then uses this mapping to make predictions. It can't handle complicated data, though, since there are non-linear correlations between the input variables and the goal variable.

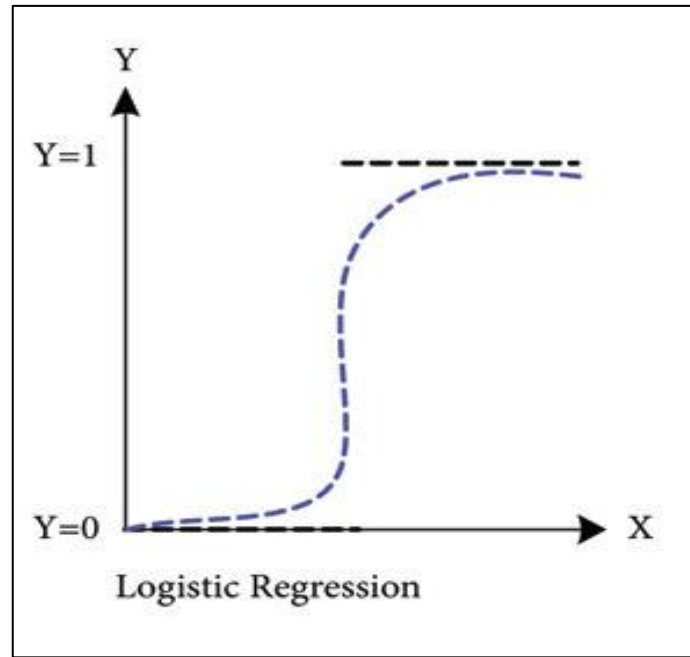


Figure 2.9: Logistic Regression [45].

2.4.5 K-Nearest Neighbors (KNN)

The k-NN model is a simple, intuitive, and non-parametric ML algorithm primarily used for classification, though it can also be applied to regression tasks [46]. Instead of developing an explicit model during the training phase, KNN classify new data works by finding the K training examples to find the nearest case in the data set, for example if ($k=1$) then simply assign the new case to the class of its first nearest neighbour so if the ($k=3$) , KNN calculate the distance of the nearest three cases and apply majority vote on the class of these cases to decide the class of the new data. The distance often calculated using metrics for finding the nearest neighbour for the numerical data is calculated by the Euclidian distance function and for the categorical data hamming distance measure.

2.4.6 Decision Tree (DT)

The DT model is a widely-used ML algorithm for both classification and regression tasks. It works by segmenting the data into subsets based on distinctive values of input features [47]. that generates the classification rule by breaking down the dataset into smaller and smaller subset until the decision node. while the root node represents the attribute with highest information gain that determines the tree branches in which each branch represents one of the outcomes of the model. These splits are determined using metrics like Gini impurity, entropy, or mean squared error, depending on the task. The end nodes, known as

leaves, represent the final predictions or decisions. Visually, a DT is easy to interpret and understand, as it mirrors human decision-making processes. One of its primary strengths is its transparent nature, which allows for clear reasoning behind predictions. However, DTs can be prone to overfitting, especially when they are deep, capturing noise and making them less generalizable.

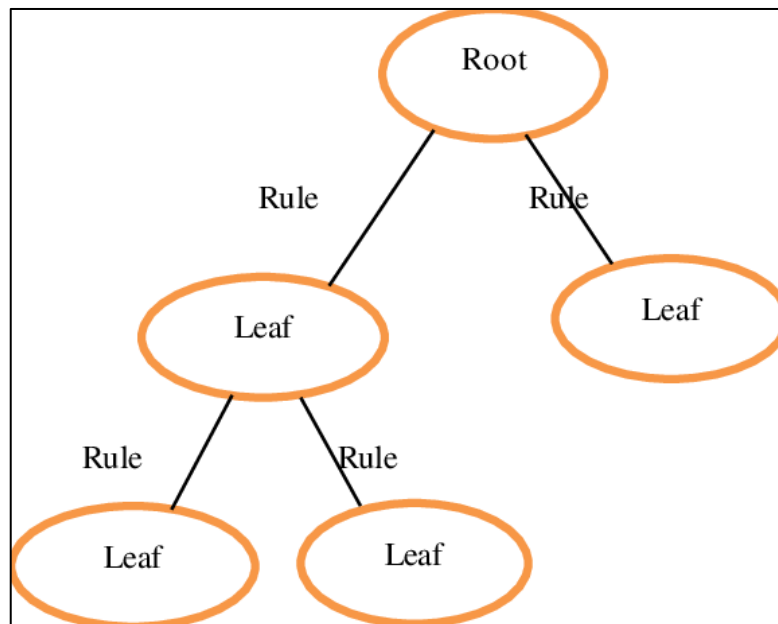


Figure 2.10: Decision Tree [48].

2.4.7 CatBoost

The model known as CatBoost, which comes from the initials of “Category Boosting,” is a GB algorithm created by Yandex with a heavy emphasis on the support of categorical features [49]. In general, GB models have serious problems that are closely related to the pre-processing and encoding of categorical variables, while CatBoost can deal with categorical features directly without needing encoding. It uses what is called ordered boosting as well as a special kind of pooling to avoid typical pitfalls faced by GB methods. An additional benefit is that this approach can effectively handle missing values and do so in an efficient manner. Moreover, CatBoost also comes with integrated visualization support that helps in evaluating the performance of the model and the relevance of features. The robustness against usual pitfalls in handling categorical data, and the ability to deliver world-class results have made CatBoost one of the sought-after players in ML competitions and applications.

2.5 EXPLAINABLE ARTIFICIAL INTELLIGENCE

The subject of explainable artificial intelligence, XAI, is considered a paradigm shift in the field of ML and AI [50]. In general advances in artificial intelligence (AI) have led to its widespread industrial adoption, with machine learning systems demonstrating superhuman performance in a significant number of tasks. However, this surge in performance, has often been achieved through increased model complexity, turning such systems into “black box” approaches and causing uncertainty regarding the way they operate and, ultimately, the way that they come to decisions. This ambiguity has made it problematic for machine learning systems to be adopted in sensitive yet critical domains, where their value could be immense, such as healthcare. As a result, scientific interest in the field of Explainable Artificial Intelligence (XAI), a field that is concerned with the development of new methods that explain and interpret machine learning models, has been tremendously reignited over recent years [51]. A user, whether they are a doctor, analyst, or ordinary person, can rely on the outputs of the model if they understand why the model chose to make certain decisions that led to those results and, in that light, diagnose and correct any flaws with greater efficiency. When it comes to numerous regulated industries, there might be a requirement – legally or ethically – to provide explanations for decisions, which means that XAI goes beyond being merely an attractive feature but becomes a necessity. It should be noted that as AI and ML are steadily penetrating different domains of our society, XAI is really important because it plays its part in ensuring an ethical and transparent deployment of AI by making sure that what is being deployed can be responsible and trustworthy.

2.5.1 Explainability and Interpretability

The topics of explainability and interpretability are usually used by researchers interchangeably; however, while these terms are very closely related, some works identify their differences and distinguish these two concepts [51]. Although these two words are sometimes used interchangeably in some cases, they represent different aspects of AI model comprehension. The concept of interpretability encompasses the idea that humans can understand the workings of a machine learning model. In this sense, interpretable models are those in which their decisions can be explained in human terms, and the procedure to derive them is also revealed. An example of an interpretable model could be a

linear regression model because its decisions are based on weighted input features, which can then be scrutinized directly and made sense out of. Accordingly, decision trees provide a readily understandable hierarchy of decisions that result in an explicit set of choices for any given situation. Understanding how a model is making its decisions can be as important as the decisions themselves; therefore, interpretability matters most when consequences are dire, such as medical diagnoses and financial predictions. It appears that explainability, however, goes even further than this into post-hoc explanations. It deals with the quality of communication between a model's internal mechanics and human users, even when these internal mechanics are not interpretable by humans themselves. This is common with many advanced models like deep learning networks, where the systems seem to work like "black boxes" with no sense of how they make their decisions. Among the goals of explainability methods is ensuring that the nature of the decision is clear and transparent, even though the way a model has arrived at it cannot be fully illuminated. Local Interpretable Model-agnostic Explanations (LIME) or SHapley Additive exPlanations (SHAP) are also instances of tools providing explanations for decisions made by sophisticated models, which shed light on different features or inputs that have had the most bearing on a model's output. In the baseline, transparency is interpreted as one in which it is easily understood by a person who sees the result how the model operates and produces output; on the other hand, explainability emphasizes providing reasons that can be comprehended by any reader of a decision taken by a model, while the complexity of the internal model does not matter. These two ideas are very important for AI and ML models that are deployed with responsibility, ethically, and effectively promoting trust and accountability of their applications in different sectors.

2.5.2 Explainable Artificial Intelligence Methods

A set of methods is designed to make the decisions of AI and ML models more transparent and explainable. These methods can be used to ensure that outputs generated by these models are explicit and interpretable, helping in building trust and ensuring accountability. In order to have a clear understanding of how this works, here are the key methods used in XAI:

- a. A. LIME is a method for approximating black-box classifiers with interpretable models on an instance-specific level [49]. Following a simple yet powerful approach, LIME can generate interpretations for single prediction scores produced by any classifier. For

any given instance and its corresponding prediction, simulated randomly-sampled data around the neighbourhood of input instance, for which the prediction was produced, are generated.

- b. One of the newer methods that has gained popularity is Shapley Additive Explanations, which takes its inspiration from cooperative game theory. It apportions a model's output – that is, the difference between the value of an instance and its mean prediction across all instances – to each input feature to measure feature importance in a unified way and assign a unique score for a feature that reflects how much it has contributed to the prediction in question.
- c. Counterfactual Explanations: This is the type of insight that answers "what if" questions. A good example is for an AI model to deny a loan application, which is a counterfactual explanation that may say: "The loan would have been approved had the applicant's income been \$10,000 more than it currently is."
- d. This is Feature Visualization – a process that involves the visualization of high-dimensional data in order to understand what features an AI model has learned. It is best utilized for neural networks where looking at the activations and filters can help someone understand which features of the input data are most significant.
- e. The Attention Mechanism in Deep Learning: Initially proposed to enhance the output of artificial intelligence neural network models, attention mechanisms can also be used to make a model more interpretable. They point out parts of the input data that are given importance by the model when coming to its conclusion.
- f. F. Rule Extraction: A simple way to do this is to turn the complex black box into a set of human-readable rules that define the decision boundary. For instance, we could extract decision trees or rule lists from neural networks so that they can provide an overview that is simpler and more interpretable regarding the model's decisions.
- g. Model-specific methods: Certain algorithms come with interpretability models because they are built into their construction. For example, decision trees are very interpretable in nature as the tree structure helps visualize the different decision pathways easily. Similarly, linear regression models provide feature coefficients that show the weight or importance of each feature.

3. METHODOLOGY

3.1 PROPOSED METHODOLOGY DESIGN

User behavior detection in the field of online advertising was our domain, and we made use of a few steps (Figure 3.1). Our first step was to collect the detailed online advertising dataset and analyze the data through observation that could lead us to insights about user behavior. This was followed by data preprocessing, which included cleaning, transforming, and getting the data ready for model training. After that, we implemented basic models like RF, GB, LR, KNN, DT, and CatBoost with or without parameter tuning. The preprocessed data were used for training purposes while the models were evaluated based on metrics.

On the other hand, in order to improve the results of the models, we incorporated a combination of soft and hard voting methods with ensemble learning techniques. The concept of soft voting is where predicted probabilities from individual models are averaged, while for hard voting, majority rule decides the final class labels. These steps were taken to develop a methodology that effectively uses ML algorithms and ensemble techniques for the detection and analysis of user behavior in online advertising. The models' performance evaluation led us to choose soft voting as our preferred method due to its highest accuracy among other models available for our data set.

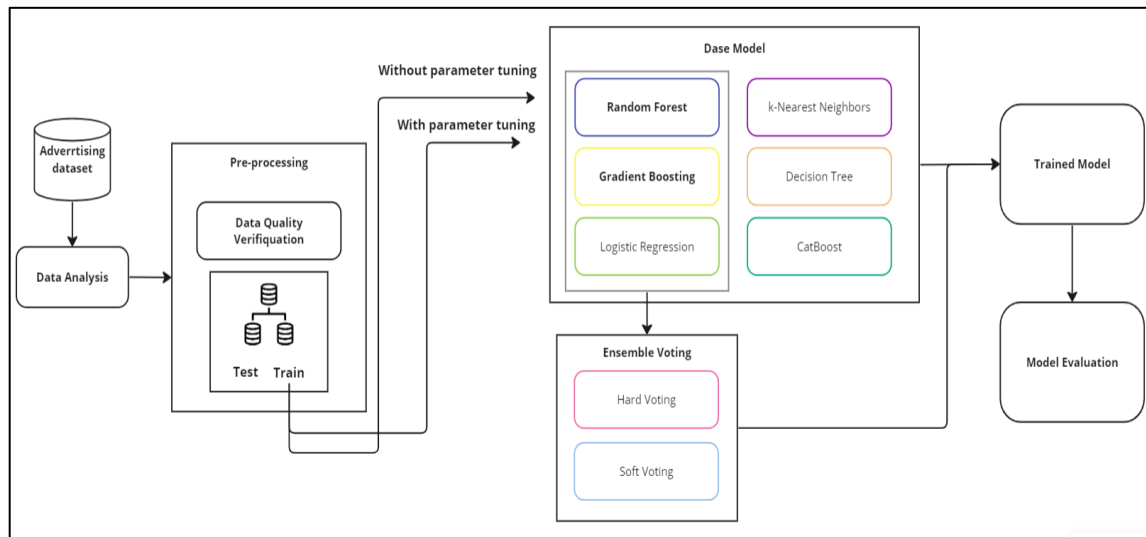


Figure 3.1: The Proposed Methodology Design.

3.1.1 Dataset Gathering

According to this work, we use a dataset of 1000 examples, each one with 8 attributes. Every example represents an individual user's experience with an online ad. The list of these features includes:

- a. Time on site is the variable under investigation, a continuous measure using minutes as the unit to describe the time spent by a user on a site every day.
- b. And an integer variable named Age should be used to store the user's age.
- c. Area Income: A continuous variable, likely representing the average income in the user's geographic area.
- d. Daily Internet Usage: Another continuous variable representing the total time a user spends on the internet daily, possibly measured in minutes.
- e. Male: A binary variable indicating the user's gender, where 1 corresponds to male and 0 corresponds to female.
- f. The label, or target variable, is "Clicked on Ad", a binary variable where '1' denotes that the user clicked on the advertisement, and '0' indicates that the user did not.

3.1.2 Data Analysis

3.1.2.1 Age distribution

The distribution of ages among our users is a critical component of understanding user interactions with online advertising. Figure 3.2 below shows a histogram illustrating this distribution.

The age of users in our dataset varies from a minimum of 19 years to a maximum of 61 years. The distribution's range helps us to understand the wide diversity in the age of users interacting with the advertisements.

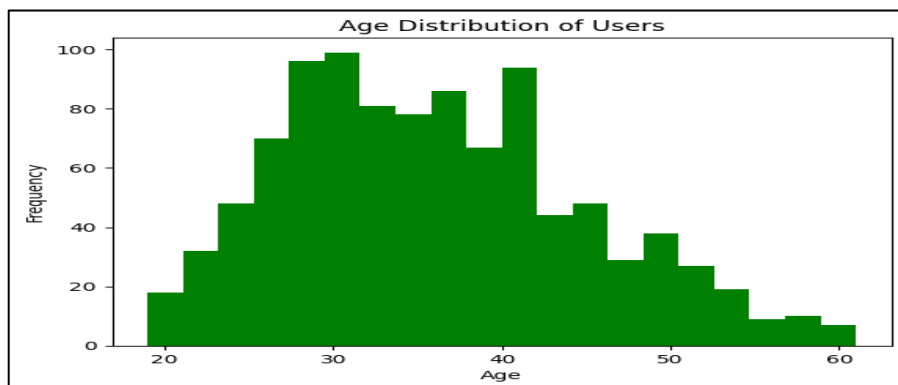


Figure 3.2: Users Age Distribution Histogram.

As we can see from the histogram, the data is bimodal. Most users interacting with the advertisement fall into the 26-32 and 38-40 age brackets. The frequency gradually increases as the age increase from 19-32, indicating a positive correlation between the age of users and their interaction with online advertisements. However, the frequency gradually decreases as the age increase from 40-61, indicating a negative correlation between the age of users and their interaction with online advertisements.

3.1.2.2 Gender distribution

The proportion of gender among the users in the system is yet another vital piece in understanding who interacts with online advertisement demographically. In our dataset, Figure 3.3 shows a bar plot illustrating the gender distribution among users.

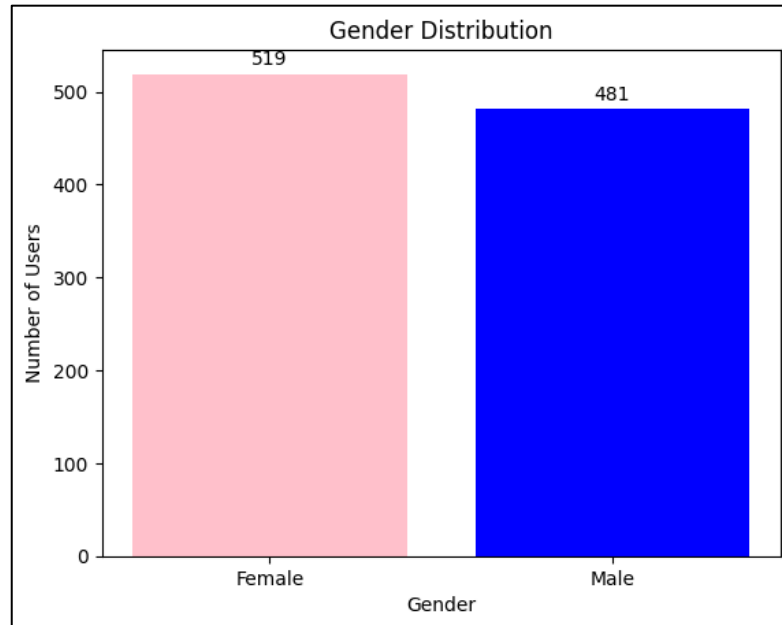


Figure 3.3: Users Gender Distribution.

Of the 1000 users we have in our database, 519 of them, or about 51.9 percent, are females and the rest, i.e. 481 of them, which is equivalent to 48.1%, are males. This relatively even gender distribution informs us that our database does not have an overt bias on either gender, meaning a good sample has been taken since a wider analysis can be done on gender patterns and influences as related to online ad interaction.

3.1.2.3 Ad effectiveness

On the other hand, another principal point to understand in our investigation is to measure how effective the ads are in front of users. This feature is represented by the attribute "Clicked on Ad," where '1' shows that a user has clicked on an advertisement while '0'

depicts that they have not. The visualization tool we use is a pie chart presented in Figure 3.4, which shows how many users click versus how many do not.

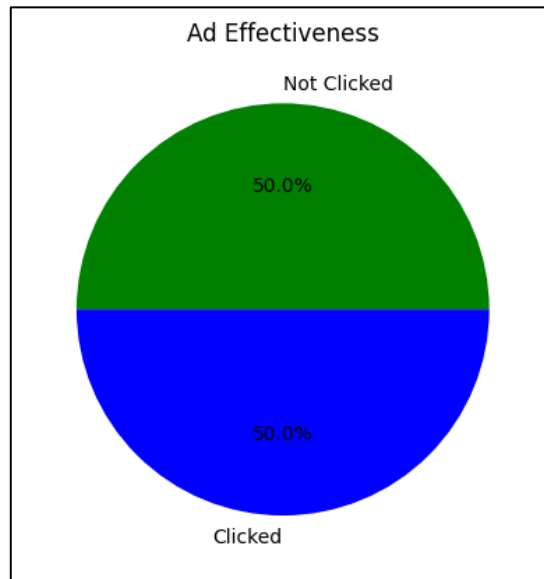


Figure 3.4: Ad Effectiveness Label Distribution.

In our dataset, we observe an equal distribution between users who clicked on the advertisements and those who did not, with both groups comprising 50% of the total users. This balanced distribution is advantageous for our study as it avoids potential bias towards one class over the other in our subsequent predictive modeling.

3.1.2.4 Correlation analysis

To understand the relationships between different attributes in our dataset, we compute a correlation matrix, which is presented in Figure 3.5.

The correlation matrix reveals several key relationships:

- a. Daily Time Spent on Site and Clicked on Ad: This pair shows a significant negative correlation of -0.748. This suggests that as the daily time spent on the site increases, the likelihood of the user clicking on an ad decreases. This might imply that users spending more time on the site are perhaps more focused on content rather than the advertisements.
- b. Age and Clicked on Ad: There's a positive correlation of 0.493, indicating that as age increases, the likelihood of clicking on an ad also increases. This could suggest that older users are more likely to interact with the advertisements.

- c. Area Income and Clicked on Ad: There's a negative correlation of -0.476. This suggests that users from areas with higher incomes are less likely to click on an ad. This could be due to a variety of factors that could be further explored.
- d. Daily Internet Usage and Clicked on Ad: The negative correlation value is -0.787, which indicates there is a strong negative correlation between people with high daily internet usage and ads clicked. This suggests that frequent internet users do not often click on advertisements, such as the feature "Daily Time Spent on Site" can tell us.
- e. Male and Clicked on Ad: The correlation is -0.038, indicating almost no linear relationship between the user's gender and clicking on an ad.

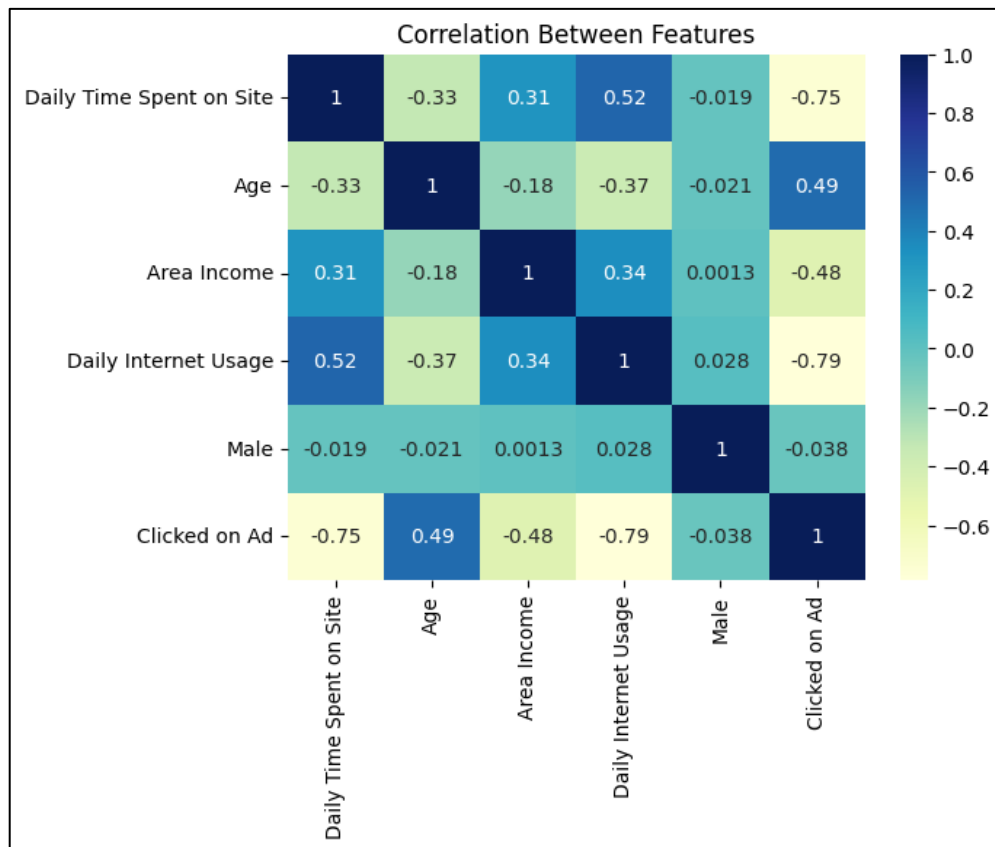


Figure 3.5: Dataset Correlation Matrix.

3.1.3 Data Pre-Processing

3.1.3.1 Data quality verification

When we are conducting a research project, one of the first things we do is to ensure the performance and consistency of our data sets. Furthermore, in order to verify the data, we used an output confirmation program on the entries so that there are no missing values or copies. Not only do we check each variable, but when performing a full check with Python

and the pandas library for whether there exist significant quantity or missing values among our data columns. Using this method we could run a complete data cleaning process over the course of this work, again avoiding anything unnecessary. We discovered that there were no missing fields (NaN) after all with our DataFrame: the `isna().sum()` column added up to zero. Furthermore, we checked our DataFrame for duplicates using the `duplicated.sum()` Function. We found that our data set has no duplicate records, so all user interaction records are unique. With no missing values and no duplicate records, this check has laid a firm foundation for all future data analyses and predictive modeling, meaning that we can be sure our findings are based on the right data. By doing so, this step forecloses upon those potential data quality problems: we might produce biased or misleading results because some data is not there or because there are two copies of it.

3.1.3.2 Data splitting

The objective is to have higher efficiency and accuracy in the data sorting process because it could better prepare datasets for subsequent predictive modeling tasks. Data slicing is of paramount importance, as it guarantees the production of a robust and trustworthy ML model. By deviating from this step, we run the risk that our model will not fit the training data well in new, unencountered data, and thereby, overfitting may occur. We have defined the feature matrix (X) and the target variable (y) in our study. The feature matrix includes 'Daily Time Spent on Site', 'Age', 'Area Income', 'Daily Internet Usage', and 'Male'. The target variable y is 'Clicked on Ad', indicating whether a user clicked on an advertisement or not.

A subset in the training set is created by using a function from the scikit-learn library that splitting operation is performed using the `train_test_split()` function. This takes the feature matrix X and the target variable y. This function shuffles the dataset and then splits it. We allocate 70% of the data to the training set and the remaining 30% to the test set. This split allows us to train our model on a large portion of the data, while still withholding so much for future tests and evaluations.

3.1.4 Proposed Classifiers

3.1.4.1 Random forest classifier

This research used the RF classifier is a machine learning algorithm that belongs to the family of ensemble methods , also ML model used in this research. RF is a method which operates by constructing a multitude of DTs at training time and outputs, at test time, the class that is the mode of the classes of the individual trees. Such an approach is, in effect, a form of crowd voting or bagging method. ML algorithms have hyperparameters, which are settings that can be tuned to control their behavior Hyper-parameter tuning is the process of finding the combination of hyperparameters for a learning algorithm that performs best for a particular problem. As to our RF classifier, the five key hyperparameters were selected to optimize tuning.

- a. “n_estimators”: signifies The number of trees in the random forest ensemble.
- b. “max_depth”: is the limit for the depth of each tree.
- c. "min_samples_split": The decision tree requires a minimum number of samples to divide an internal node. It helps to control the complexity and overfitting of the tree.
- d. "min_samples_leaf": The minimum number of samples can a leaf node contain?
- e. "max_features": How many features of the optimal split for each node do you want to take into account? We entered "auto," "sqrt," or "log2" or through an integer. Reducing the inter-between-tree association and increasing the variability of the forest are can be accomplished by specifying a smaller number .

We used Grid Search to find the optimal hyperparameters.This is a hyperparameter optimization method that involves combing through a specific set of variables. We chose several hyperparameters based on their significance effects. We chose several hyperparameters based on their significance as well as the known effect in this learning method. For each parameter we selected a range of possible values, and the Grid Search method measured the RF model Performance in the different combinations of these hyper.

```
param_grid = {  
    'n_estimators': [50, 100, 200],# Number of trees in the forest  
    'max_depth': [None, 10, 20, 30],# Maximum depth of the tree  
    'min_samples_split': [2, 5, 10],# Minimum number of samples required to split an internal node  
    'min_samples_leaf': [1, 2, 4], # Minimum number of samples required to be at a leaf node  
    'max_features': ['auto', 'sqrt', 'log2'], # Number of features to consider when looking for the best split  
}
```

Figure 3.6: Random Forest Parameters Grid.

With the RF classifier, the best hyperparameters {'max_depth': 20, 'max_features': 'log2', 'min_samples_leaf': 4, 'min_samples_split': 2, 'n_estimators': 50} were then used to train an ensemble model for the final result.

3.1.4.2 Gradient boosting classifier

In our investigation, we also made use of the GB classifier, which is an ML model. A group of weak prediction models, usually DTs, are combined to construct a prediction model using the GB EL technique.

By permitting the optimization of any differentiable loss function, it expands on the model's step-by-step construction.

In this context, we focused on tuning six key hyperparameters: `n_estimators` and `learning_rate`.

- a. `n_estimators` is a representation of the total number of boosting steps needed. A little number could result in underfitting, while a large number could cause overfitting.
- b. Each tree's contribution to the ensemble is regulated by `learning_rate`. To model every relation, a lower learning rate needs more trees, but the predictions are frequently more accurate.
- c. "`max_depth`": This variable indicates the maximum depth that each distinct regression estimator is permitted to have. While increasing '`max_depth`' may result in a more complicated model, it also raises the possibility of overfitting. Considering the intricacy of the issue and the information at hand, think about testing out several numbers.
- d. "`min_samples_split`": When constructing a tree, this variable determines the absolute minimum number of samples required to split an internal node. Higher values prevent overfitting by increasing the number of samples needed for a split.
- e. "`min_samples_leaf`": This indicates the bare minimum of samples needed at a leaf node. Higher values can aid in preventing overfitting and controlling the complexity of the separate regression estimators.
- f. "`subsample`": This denotes the percentage of samples that will be utilized to fit each unique tree. A portion of The training data is used to train the model using any of the

learning techniques for values less than 1.0. This can improve the model's generalization to new data and lessen overfitting.

We implemented Grid Search to optimize these hyperparameters. Grid Search involves exhaustively considering all parameter combinations and retaining the best combination that gives the highest performance, based on a specified scoring metric.

```
param_grid = {  
    'n_estimators': [50, 100, 200], # Number of boosting stages (trees) to perform  
    'learning_rate': [0.01, 0.1, 1], # Learning rate shrinks the contribution of each tree  
    'max_depth': [3, 5, 7], # Maximum depth of the individual regression estimators  
    'min_samples_split': [2, 4, 6], # Minimum number of samples required to split an internal node  
    'min_samples_leaf': [1, 2, 3], # Minimum number of samples required to be at a leaf node  
    'subsample': [0.6, 0.8, 1.0], # Fraction of samples used for fitting the individual trees  
}
```

Figure 3.7: Gradient Boosting Parameters Grid.

The optimal hyperparameters for the GB classifier were determined to be { 'learning_rate': 1, 'max_depth': 3, 'min_samples_leaf': 2, 'min_samples_split': 6, 'n_estimators': 100, 'subsample': 0.8}. These parameters were then used to train the model for the final ensemble.

3.1.4.3 Logistic regression classifier

LR is a popular SL algorithm used for binary classification problems. It is a linear model that uses a logistic function to model the relationship between the input features and the probability of belonging to a specific class. In LR, the input features are combined linearly using weights, and then passed through a logistic (sigmoid) function that maps the linear combination to a probability between 0 and 1. The probability denotes the chance of the sample falling into the positive category. The sample is categorized as belonging to the positive class (clicked) if the probability is higher than a predetermined threshold, which is usually 0.5; if not, it is categorized as belonging to the negative class (non-clicked). Here, we concentrated on adjusting the following five crucial hyperparameters:

- a. "C" and "penalty." 'C' represents the inverse of the regularization strength. It controls the trade-off between fitting the training data well and avoiding overfitting. Smaller values of 'C' indicate stronger regularization, while larger values indicate weaker regularization.

- b. 'penalty' specifies the type of regularization used in the LR model. 'l1' refers to L1 regularization, also known as Lasso regularization, which encourages sparsity by shrinking some coefficients to exactly zero. 'l2' refers to L2 regularization, also known as Ridge regularization, which shrinks the coefficients towards zero without making them exactly zero.
- c. 'fit_intercept': Will decide whether to have an intercept term in the LR model. Set to True, it will learn an intercept term; if False, no intercept will be taken into account. Whether to include an intercept should depend on the problem and on the data.
- d. 'solver': This specifies what algorithm to use for optimization. Python's scikit-learn has various own solvers for LR, each with its own characteristics and strengths. 'liblinear' and 'saga' are two of the more popular. 'Liblinear' is good for smaller and medium-sized datasets, while 'saga' is better suited to large datasets.
- e. 'max_iter': The parameter represents the solver needs more than this number of iterations to reach a constraint. An optimization algorithm is solving a problem and arrives at the optimal solution when it has found one. If the solver can't reach within the number of retries allowed, it is possible that an optimal solution has not been found. You can confine the parameter based on the convergence behavior of the model.

By trying out different combinations of the various parameters, we can find the combination that produces the best model performance. This process is quite helpful in discovering hyperparameter values that generalize well to new data and enhance the performance of this model.

We implemented Grid Search to optimize these hyperparameters.

```
param_grid = {'C': [0.001, 0.01, 0.1, 1, 10], # Inverse of regularization strength
'penalty': ['l1', 'l2'], # Regularization type
'fit_intercept': [True, False], # Whether to include an intercept term in the model
'solver': ['liblinear', 'saga'], # Algorithm to use for optimization
'max_iter': [100, 200, 500], # Maximum number of iterations for the solver# Add more parameters here as needed
}
```

Figure 3.8: Logistic Regression Parameters Grid.

Optimal hyperparameters for the GB classifier were determined to be:

{ 'C': 1, 'fit_intercept': True, 'max_iter': 100, 'penalty': 'l1', 'solver': 'liblinear'}. These were then used to train the model for the final ensemble.

3.1.4.4 K-nearest neighbors classifier

The KNN classifier, a non-parametric and instance-based learning algorithm, was explored. The primary mechanism of KNN involves identifying the "k" closest data points (or neighbors) from the training set for each test point and assigning the most frequent label among these neighbors as the predicted class.

To ensure the most optimized performance of the k-NN classifier, a systematic hyperparameter tuning was carried out.

The parameters under consideration included:

- a. `n_neighbors`: The number of neighbors to be taken into account, with tested values being 1, 3, 5, and 7.
- b. `weights`: The strategy for assigning weights to the neighbors, evaluated under two schemes - 'uniform' (where all neighbors are given equal weight) and 'distance' (where closer neighbors are given more weight).
- c. `p`: The power parameter for the Minkowski distance metric, determining the type of distance calculation. A value of 1 corresponds to the Manhattan distance, while a value of 2 represents the Euclidean distance.

The GridSearchCV method was used to take all the combinations of these hyperparameters into account and find the most suitable configuration. This method exhaustively computed over the specified parameter values, and the hyperparameters at last used in the final model were those that yielded the best performance on the training data, with a 3-fold cross-validation mode.

```
# Define the parameter grid for KNN
param_grid = {
    'n_neighbors': [1, 3, 5, 7], # Number of neighbors to consider
    'weights': ['uniform', 'distance'], # Weighting scheme for neighbors
    'p': [1, 2] # Power parameter for the Minkowski distance metric (1 for Manhattan, 2 for Euclidean)
}
```

Figure 3.6: KNN Parameters Grid.

Upon completion of this grid search, we found out the k-NN classifier should be configured {'n_neighbors': 5, 'p': 1, 'weights': 'distance'}. We developed our model with it in mind. For many problems, taking 5 neighbors, using Manhattan distance (p=1), and assigning weights based on distance works best, and this problem was no exception.

3.1.4.5 Decision tree classifier

The DT is a flowchart-based structure in which internal nodes signify traits, branches represent decisions, and leaf nodes are the outcomes or categories. When examining the classification result, the algorithm is based on questions, and following the path that corresponds to the answer, it can make a "decision".

To obtain the best performance from the DT classifier and optimize it for the data and problem in question, a comprehensive process for tuning the hyperparameters was conducted. What hyperparameters were considered as factors for optimization:

- a. criterion: The function used to measure the quality of a split. The two parameters it included were 'gini' for Gini impurity and 'entropy' for information gain.
- b. max_depth: The maximum depth of the tree. It has already various values, including None (indicating no limit), 10, 20, or 30.
- c. min_samples_split: Specifies the minimum number of samples able to split an internal node. By trying values of threshold 2, 5, and 10 no impurity was left in the parent node.
- d. min_samples_leaf: Indicates the lowest limit on samples to use for a leaf node. The choices are 1, 2, and 4.

The hyperparameters can be optimally combined using the GridSearchCV method. This method works by conducting an exhaustive search over the specified hyperparameters to determine not only which combination of them offers the best performance, but also whether that is enough and validated using a 3-fold cross-validation strategy. After fitting the DT model with different parameter combinations on the training data, GridSearchCV makes sure that hyperparameters deliver best predictive performance for any given problem.

```

param_grid = {
    'criterion': ['gini', 'entropy'], # Splitting criterion
    'max_depth': [None, 10, 20, 30], # Maximum depth of the tree
    'min_samples_split': [2, 5, 10], # Minimum samples required to split an internal node
    'min_samples_leaf': [1, 2, 4] # Minimum samples required at a leaf node
}

```

Figure 3.9: DT Parameters Grid.

The optimal hyperparameters for the DT classifier were determined to be:

{'criterion': 'gini', 'max_depth': 30, 'min_samples_leaf': 2, 'min_samples_split': 10}

3.1.4.6 CatBoost classifier

In this paper, the CatBoost classifier, an advanced GB algorithm is explored. CatBoost, which emerged from Yandex, is special because so successful because it natively supports categorical data, dispensing with the need for extensive pre-processing; or, much less so, time-consuming manual encoding, as does performing these tasks separately using many other ML algorithms.

The objective of fine-tuning the CatBoost classifier so that it fits well with the specific data and problem of this research, is to do Hyperparameter Optimization.

The considered hyperparameters encompassed:

- iterations: Refers to the number of boosting stages the algorithm should run, with values explored being 100, 200, and 300.
- learning_rate: Dictates the speed of the model's training, with candidate values being 0.01, 0.1, and 0.2.
- depth: Represents the depth of the trees in the model, and the study evaluated depths of 4, 6, and 8.
- l2_leaf_reg: An L2 regularization term on weights, introduced to reduce overfitting, with tested coefficients being 1, 3, and 5.

To ascertain the most optimal combination of these hyperparameters, the GridSearchCV method was employed. This method performs a comprehensive search over the provided hyperparameter values to identify the combination that results in the best performance. Using a 3-fold cross-validation strategy, GridSearchCV validates the efficacy of each combination, ensuring the selected hyperparameters are the most suitable for the task at hand.

```
# Define the parameter grid for CatBoostClassifier
param_grid = {
    'iterations': [100, 200, 300],      # Number of boosting iterations
    'learning_rate': [0.01, 0.1, 0.2],  # Learning rate
    'depth': [4, 6, 8],                 # Depth of the tree
    'l2_leaf_reg': [1, 3, 5],           # L2 regularization coefficient
}
```

Figure 3.10: Catboost Parameters Grid.

The optimal hyperparameters of the CatBoost classifier are determined to be:

(depth=4, iterations=300, l2_leaf_reg=3, learning rate=0.01)

3.1.4.7 Ensemble model

After fitting the basic models and tuning their parameters on the data, we further improved our analysis with ensemble voting methods, mainly soft voting and hard voting. What we sought to achieve using ensemble voting was the collective wisdom of individual models and whether these techniques could be more effective.

To generate the final prediction using soft voting, the predicted probabilities of the individual models are aggregated. Such a method takes into consideration the confidence level of predictions from each model. so that the decision-making process can be made more nuanced and probabalistic. On the other hand, hard voting, which combines the class votes of the individual models collapses down to the final prediction. This approach simplifies the decision-making process by considering only the most common prediction.

We hoped to use soft and hard voting to see if our finely tuned models brought any improvement in forecast accuracy or performance. We wanted to see whether this voting process made any improvement in codifying the effectiveness of service businesses vis-a-vis brand confidence.

4. RESULTS AND DISCUSSION

4.1 THE USED LIBRARIES

We made extensive use of various libraries in order to facilitate our study and implementation of machine learning models. For example, the following:

- a. Pandas is a great library for manipulating and analyzing data. It is really helpful for people who understand problem data as a table. It provides convenient data structures and functions for working with structured data efficiently, such as data frames. We loaded and preprocessed our dataset in Pandas, ran tests and performed data exploration. Then, we used it to divide our set Accordingly into training and testing datasets.
- b. NumPy is an essential library for Python numerical computations and provides a range of mathematical functions and multi-dimensional array support. We used NumPy to compute figure operations and transform data, such as scaling and array manipulation that we later used to train models.
- c. Moreover, Matplotlib is a popular library for data visualization in Python. It provides a wide variety of plotting functions and customizability options. We made use of Matplotlib to create visualizations of all kinds, including line plots, scatter diagrams, and bar charts that would allow for a better grasp of the data and present the results of our analysis.
- d. Seaborn, a high-level data visualization library. With its more pleasingly streamlined interface, you can use Seaborn to generate statistical graphics. We have made use of Seaborn to create some extremely "readable" and easy-on-the-eyes visualizations. For example, we generated heatmaps and distribution plots to investigate relationships between variables and their distributions.
- e. P.Sk-learn is a widely used ML library in Python. It is supplied with a complete set of tools for dealing with all ML-related tasks: to select models or preprocess data. Our dataset was partitioned into two parts by Scikit-learn's `train_test_split` function: one training set subsidiary and one testing set. This meant we split off some of the data to teach the models using known inputs and can test the system against unknown inputs.

4.2 EVALUATION METRICS

Compared to the previous way we evaluated models, using an interdisciplinary analysis of diverse indicators broadened our vision and taught us new things about them. Our research utilized 3 cross validations to fine-tune hyperparameters in ML models with the grid search method. By using a cross-validation of 3 to perform grid search we also ensured that the performance of the models was evaluated through a rigorous and reliable process. The above metrics are all explained below. TP is the total number of true positives, TN the number of true negatives, FP the number of false positives, and FN the number of false negatives. Compared to the previous way we evaluated models, using an interdisciplinary analysis of diverse indicators broadened our vision and taught us new things about them. Our research utilized 3 cross validations to fine-tune hyperparameters in ML models with the grid search method. By using a cross-validation of 3 to perform grid search we also ensured that the performance of the models was evaluated through a rigorous and reliable process. The above metrics are all explained below. TP is the total number of true positives, TN the number of true negatives, FP the number of false positives, and FN the number of false negatives.

$$ACC = (TP + TN) / (TP + TN + FP + FN)$$

Precision (PRE): This is the proportion of correctly predicted positive instances among all the instances labeled as positive by the model. Therefore, it shows how right the model is in positive prediction and how well it can reduce false positive errors at the same time.

$$PRE = TP / (TP + FP)$$

The recall of a model is the percentage of positive cases it can accurately identify. Among the number of real positive instances, what is the accurate number of positive examples predicted.

$$REC = TP / (TP + FN)$$

F1-score (F1-S): is the harmonic mean sensitivity of PRE and REC. It provides a balanced measure of the model's performance. F1-score is the harmonic mean of precision and recall, providing a balanced measure of the model's performance. For most target variables, the RFC algorithm achieved higher F1-scores, indicating a good balance between precision and recall. Different algorithms performed better or worse depending on the target.

$$F1-S = 2 * (PRE * REC) / (PRE + REC)$$

4.3 MODELS PERFORMANCE WITHOUT PARAMETER TUNING

4.3.1 Evaluation of the Random Forest Model

The positive predictions made by the RF model were correct. The Accuracy of 97.3% suggests without parameter tuning, that the RF technique accurately identified 97.3% of the positive samples, indicating a high level of overall correctness in its predictions. The model's performance can be further analyzed using the CM and the CR. The CM reveals that out of the total samples, 141 were correctly classified as belonging to class 0, while 151 were correctly classified as belonging to class 1. However, there were 5 instances where the model incorrectly predicted samples as class 1 when they actually belonged to class 0, and 3 instances where samples belonging to class 1 were incorrectly predicted as class 0. The CR provides an assessment of PRE, REC, and F1-S for each class.

The CR for the RF model reveals its strong performance for both classes. With a PRE of 0.98 for Class 0 and 0.97 for Class 1, the model accurately identified the majority of instances for each class. Additionally, the REC values of 0.97 for Class 0 and 0.98 for Class 1 demonstrate the model's ability to effectively capture positive instances. The balanced F1-Ss of 0.97 for both classes indicate a good trade-off between PRE and REC.

It achieved a high sensitivity, indicating that it correctly classified 98.1% of the positive instances. On the other hand, it achieved a specificity of 96.6%, indicating that it correctly classified 96.6% of the negative instances.

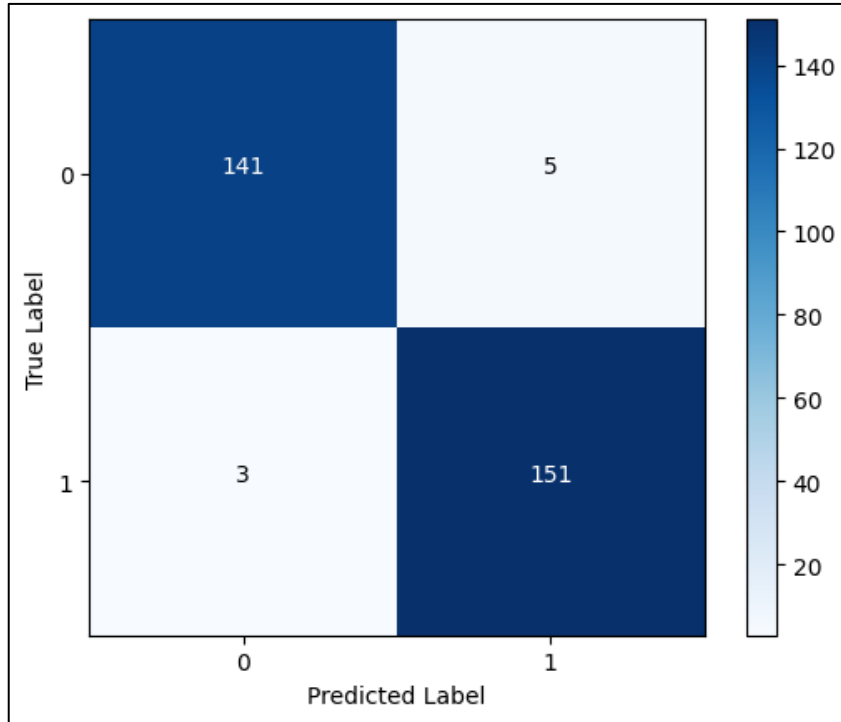


Figure 4.1: Random Forest Confusion Matrix.

Classification Report:				
	precision	recall	f1-score	support
0	0.98	0.97	0.97	146
1	0.97	0.98	0.97	154
accuracy			0.97	300
macro avg	0.97	0.97	0.97	300
weighted avg	0.97	0.97	0.97	300

Figure 4.2: Random Forest Classification Report.

Figure 4.3 illustrates the feature importance as determined by the LIME method when applied to a RF model. The figure highlights the significance of various features and their corresponding weight ranges in influencing the model's decisions. "Daily Internet Usage" within the range of 185.45 to 220.06 has a negative weight, indicating its potential in decreasing the likelihood of the predicted outcome. On the other hand, features like "Age" greater than 41, "Area income" less than or equal to 47,332.82, "Daily Time Spent on Site" between 51.47 and 68.22, and "Male" value less than or equal to 0 showcase positive weights, implying their roles in increasing the probability of the model's prediction. The magnitude of these weights provides insights into the relative importance of each feature in the model's decision-making process.

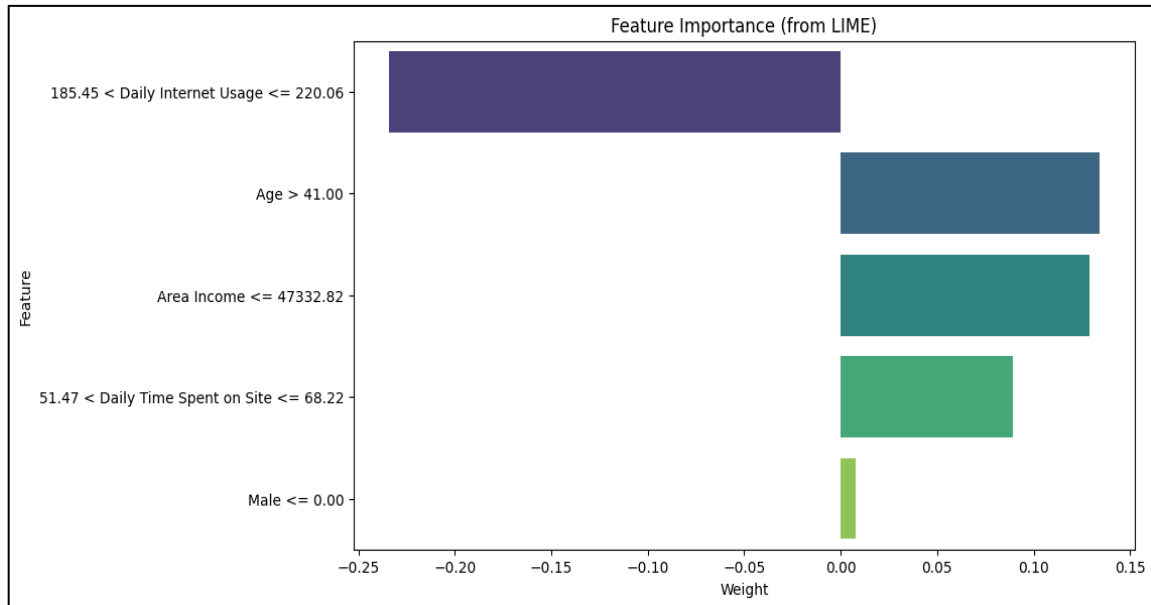


Figure 4.3: Feature Importance Derived from LIME Analysis on a RF Model.

Figure 4.4 displays a vertical bar chart that visualizes the average impact on model output magnitudes derived from SHAP for the RF method, broken down by two distinct classes. Each feature's influence is presented by a pair of bars side by side, with Class 1 represented in blue and Class 0 in red. For "Daily Internet Usage," Class 1 has an influence ranging from 0 to 0.22, followed immediately by Class 0's impact extending from 0.22 to 0.45. The "Daily Time Spent on Site" showcases Class 1's range from 0 to 0.18, and Class 0's influence from 0.18 to 0.38. When observing "Area Income," Class 1's impact spans from 0 to 0.05, with Class 0 following from 0.05 to 0.11. For the "Age" feature, Class 1's influence ranges between 0 to 0.05, slightly overlapping with Class 0's range of 0.04 to 0.1. Lastly, the "Male" feature demonstrates minimal variations, with Class 1's range from 0 to 0.001 and Class 0's from 0.001 to 0.0015. The bar chart succinctly highlights the differences in feature impact for the two classes, providing a visual comparative analysis.

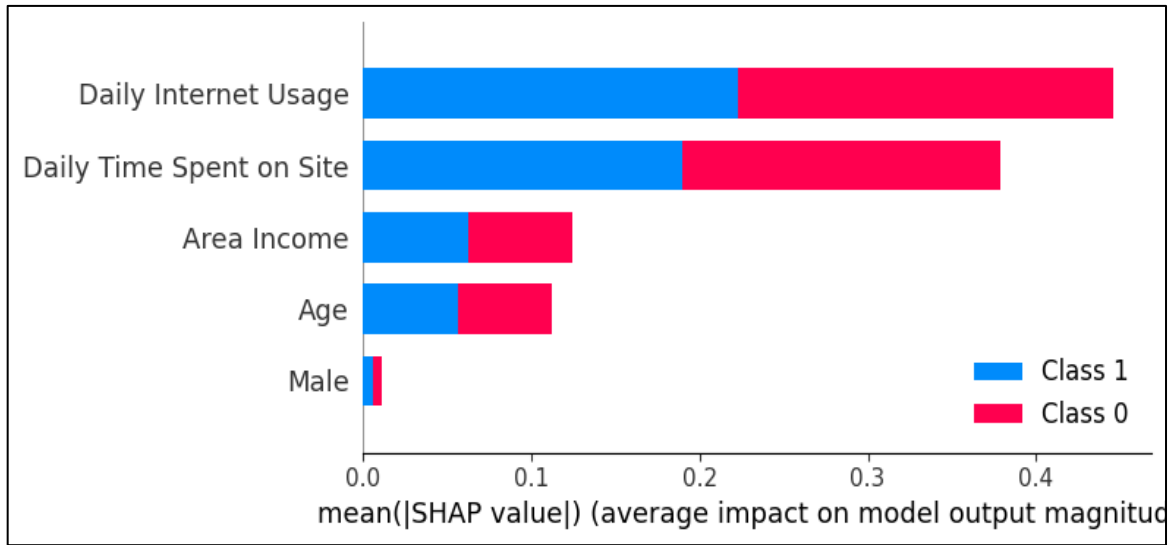


Figure 4.4: Average Impact on Model Output Magnitudes for RF Using SHAP.

4.3.2 Evaluation of the Logistic Regression Model

The LR model without parameter tuning achieved an ACC of 0.967, indicating a high level of overall correctness in its predictions. The CM provides additional insights into the model's performance. Out of the total samples, 142 were correctly classified as belonging to class 0, while 148 were correctly classified as belonging to class 1. However, there were 4 instances where the model incorrectly predicted samples as class 1 when they actually belonged to class 0, and 6 instances where samples belonging to class 1 were incorrectly predicted as class 0.

The CR for the LR model without parameter tuning demonstrates its strong performance for both classes. With a PRE of 0.96 for Class 0 and 0.97 for Class 1, the model accurately identified the majority of instances for each class. Additionally, the REC values of 0.97 for Class 0 and 0.96 for Class 1 highlight the model's ability to effectively capture positive instances. The balanced F1-S of 0.97 for both classes indicate a good balance between PRE and REC.

It achieved a sensitivity of 96.1%, indicating that it correctly classified 96.1% of the positive instances and achieved a specificity of 97.3%, indicating that it correctly classified 97.3% of the negative instances.

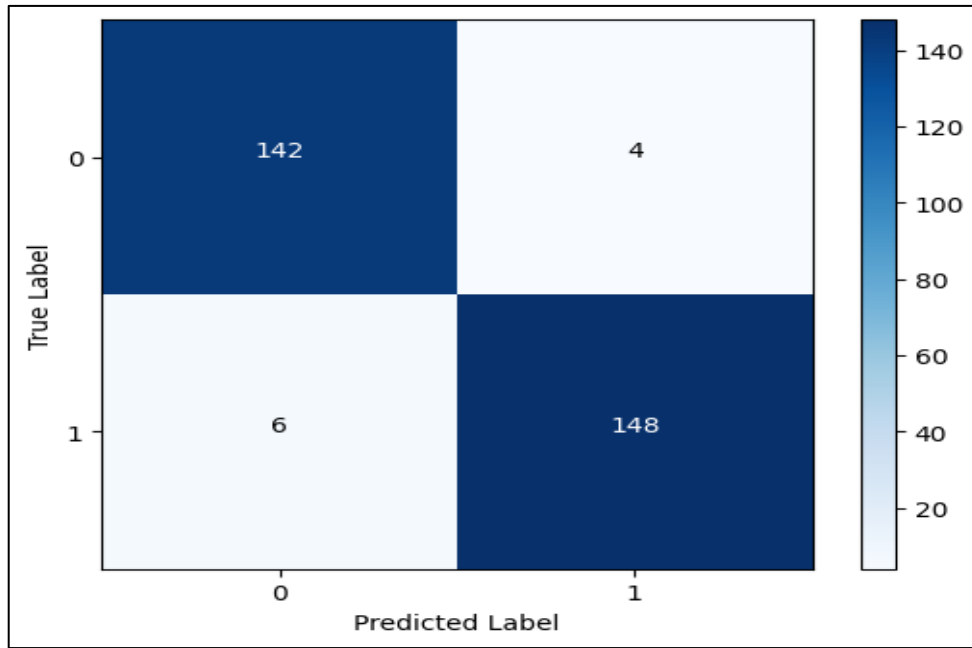


Figure 4.5: Logistic Regression Confusion Matrix.

Classification Report:				
	precision	recall	f1-score	support
0	0.96	0.97	0.97	146
1	0.97	0.96	0.97	154
accuracy			0.97	300
macro avg	0.97	0.97	0.97	300
weighted avg	0.97	0.97	0.97	300

Figure 4.6: Logistic Regression Classification Report.

Figure 4.4 showcases the feature importance as ascertained by the LIME method for a LR model. This visualization delineates the impact of specific features on the model's predictions by highlighting their associated weight ranges. The "Daily Internet Usage" between 185.45 and 220.06, "Age" greater than 41, and "Daily Time Spent on Site" between 51.47 and 68.22 all possess positive weights, suggesting their influence in amplifying the model's predictive outcome. Conversely, "Area income" less than or equal to 47,332.82 has a negative weight, indicating its role in reducing the likelihood of the predicted result. The "Male" feature with a value less than or equal to 0 exhibits a slight negative weight, hinting at its minor contribution in steering the model's decision. This figure provides a concise view of feature significance within the context of a LR model's decision-making process.

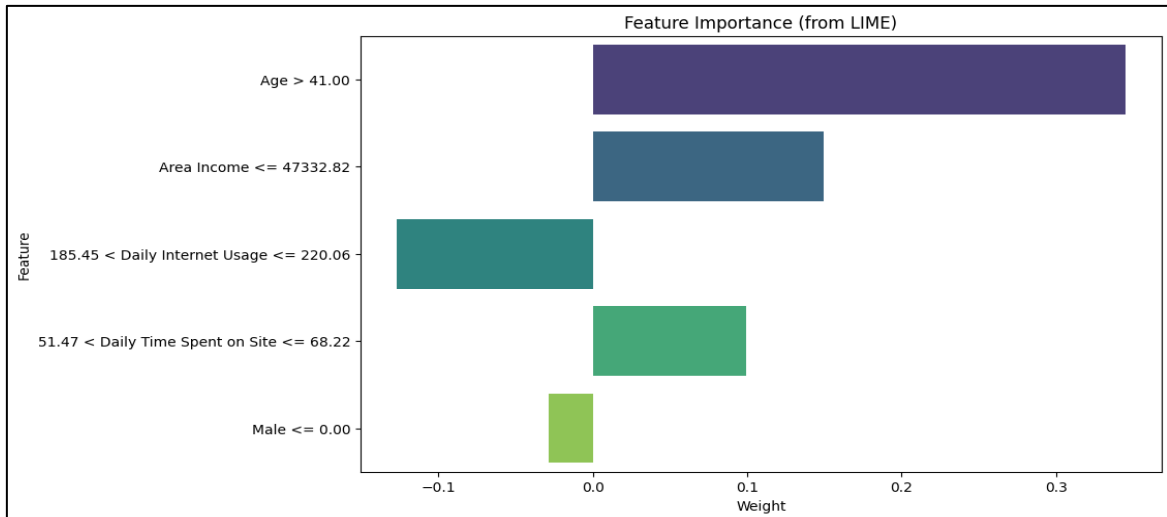


Figure 4.7: Feature Importance Derived from LIME Analysis on a LR Model.

Figure 4.12 illustrates a vertical bar chart that represents the average impact on model output magnitudes derived from SHAP when employing the LR method. In the figure, Class 1 is distinctly represented in blue. The influence of "Daily Internet Usage" for Class 1 stretches from 0 to 0.28. The "Age" feature for Class 1 showcases a notable range, extending from 0 to 2.5. Similarly, the "Daily Time Spent on Site" also holds an impact range of 0 to 2.5 for Class 1. "Area Income" presents a more moderate influence for Class 1, spanning from 0 to 1.0. Lastly, the "Male" feature demonstrates a range of 0 to 0.2 for Class 1. The bar chart offers a visual representation of the features' impacts on the model's output for Class 1, emphasizing their respective magnitudes.

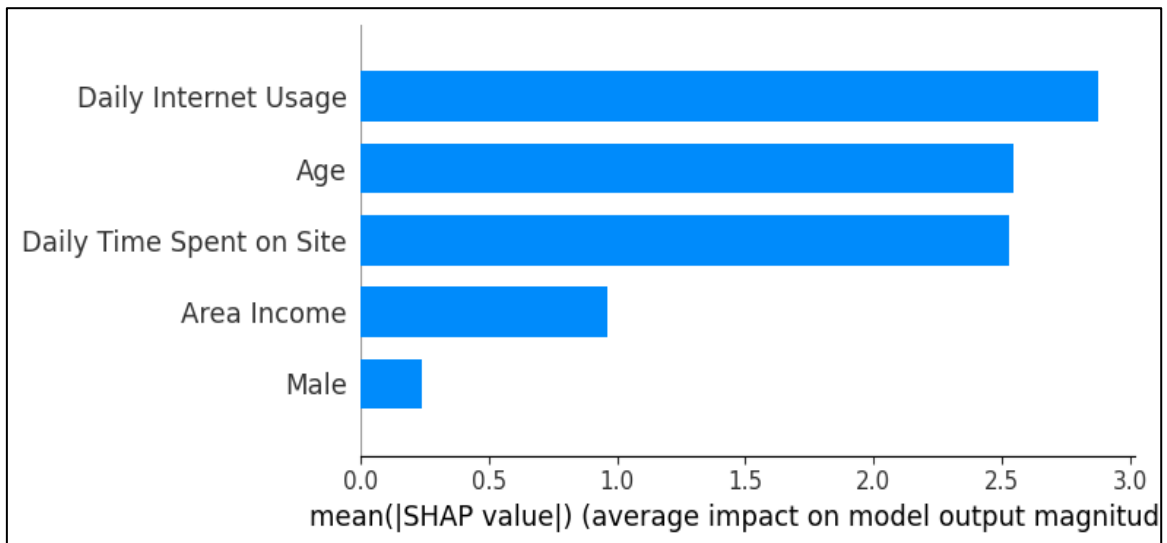


Figure 4.8: Average Impact on Model Output Magnitudes for LR using SHAP.

4.3.3 Evaluation of the Gradient Boosting Model

The ACC achieved by the model is 96.7%. The CM for the GB model without parameter tuning provides insights into its performance. Among the total samples, 138 were accurately classified as class 0, and 152 were correctly classified as class 1. However, there were 8 instances where the model misclassified samples as class 1 when they actually belonged to class 0, and 2 instances where samples from class 1 were mistakenly classified as class 0.

The CR for the GB model without parameter tuning reveals strong performance for both classes. For Class 0, the model achieved a PRE of 0.99, indicating that 99% of the instances predicted as Class 0 were correctly classified. The REC for Class 0 is 0.95, indicating that the model identified 95% of the instances belonging to Class 0 correctly. The F1-score for Class 0 is 0.97, signifying a good balance between correctly identifying Class 0 instances and minimizing false positives. For Class 1, the model achieved a PRE of 0.95, indicating that 95% of the instances predicted as Class 1 were correctly classified. The REC for Class 1 is 0.99, meaning that the model identified 99% of the instances belonging to Class 1 correctly. The F1-score for Class 1 is 0.97, suggesting a good balance between correctly identifying Class 1 instances and minimizing false positives. Results also demonstrate a sensitivity of 0.987 and a specificity of 0.945.

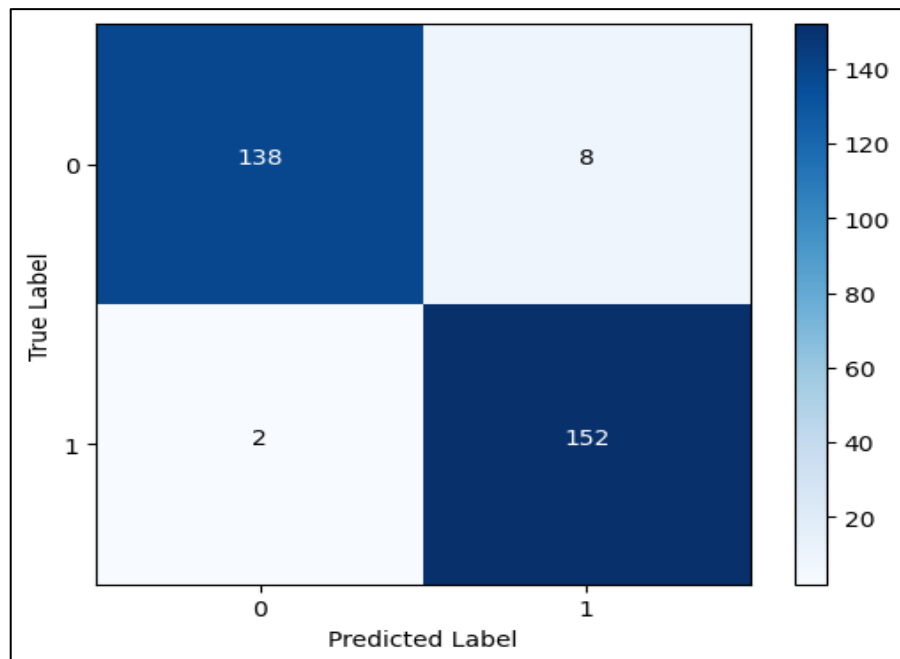


Figure 4.9: Gradient Boosting Confusion Matrix.

Classification Report:				
	precision	recall	f1-score	support
0	0.99	0.95	0.97	146
1	0.95	0.99	0.97	154
accuracy			0.97	300
macro avg	0.97	0.97	0.97	300
weighted avg	0.97	0.97	0.97	300

Figure 4.10: Gradient Boosting Classification Report.

Figure 4.11 displays the feature importance as deduced from LIME analysis when applied to a GB model. The figure underscores the influence of distinct features in shaping the model's decisions by indicating their respective weight intervals. The "Daily Internet Usage" between 185.45 and 220.06 holds a negative weight, hinting at its role in diminishing the predicted outcome's likelihood. Conversely, the features "Age" exceeding 41, "Area income" not surpassing 47,332.82, "Daily Time Spent on Site" spanning from 51.47 to 68.22, and "Male" equaling 0 or less, all exhibit positive weights, signifying their contribution to increasing the probability of the model's prediction. The magnitude of these weights provides a snapshot of the relative importance each feature holds within the GB model's decision framework.

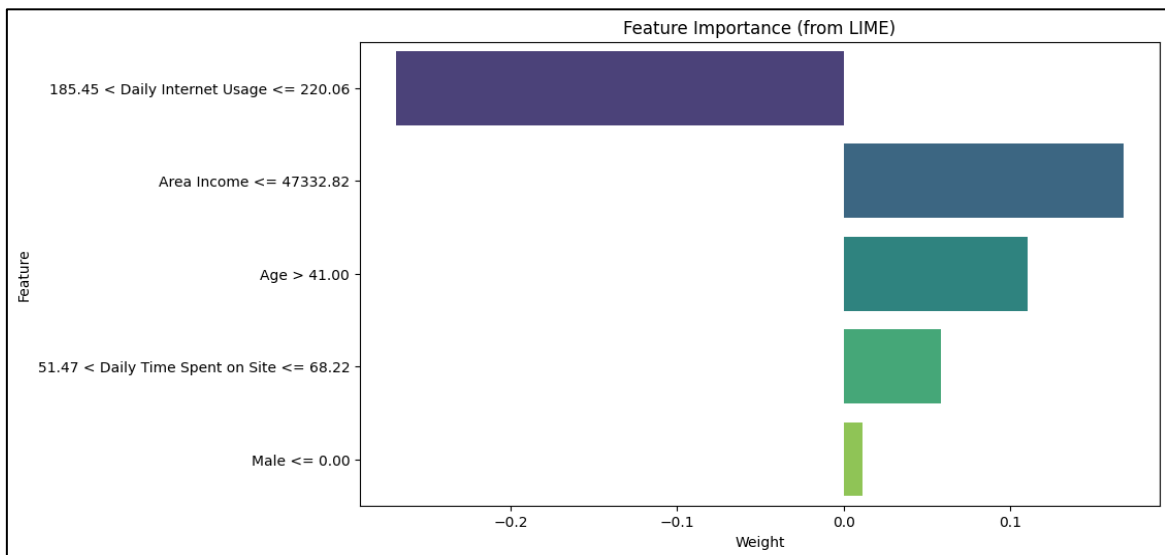


Figure 4.11: Feature Importance Derived from LIME Analysis on a GB Model.

Figure 4.12 displays a vertical bar chart showcasing the average impact on model output magnitudes using the SHAP values for the GB method. In this visual representation, Class 1 is distinctly marked in blue. The "Daily Internet Usage" for Class 1 ranges from 0 to

0.25, indicating its influence on the model's output. For "Daily Time Spent on Site", the impact spans from 0 to 2.0 for Class 1. "Area Income" demonstrates an influence ranging from 0 to 0.8 for Class 1. The "Age" feature, meanwhile, has an impact stretching from 0 to 0.5 for Class 1. Lastly, the "Male" attribute holds a more contained range, from 0 to 0.1, for Class 1. The chart succinctly captures the significance of each feature's impact on the model's predictions for Class 1 when applying the GB technique.

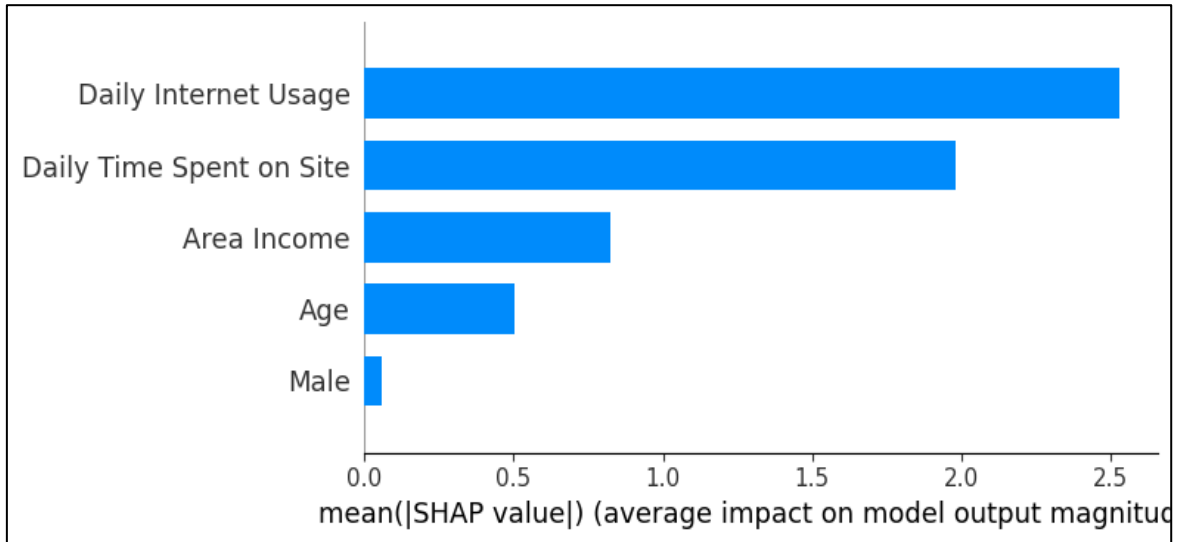


Figure 4.12: Average Impact on Model Output Magnitudes for GB Using SHAP.

4.3.4 Evaluation of the KNN Model

In the evaluation of the KNN model, an ACC of 71% was observed, suggesting a moderate level of effectiveness in classification. To gain further insights into the model's performance, the CM was examined. The matrix revealed that 109 samples were accurately classified as class 0, and 104 were correctly identified as class 1. However, there was a notable number of misclassifications: 37 instances were incorrectly categorized as class 1 when they were actually in class 0, and 50 instances of class 1 were misidentified as class 0. While the model demonstrated a fair level of ACC, these misclassifications indicate areas for potential improvement to enhance its predictive PRE and REC, ensuring a more balanced and reliable classification performance.

Upon examining the CR for the KNN model, a detailed picture of its performance metrics emerges. The model boasts an overall ACC of 83%, indicating a strong degree of reliability in its predictions. When evaluating class-specific metrics, for class 0, the model achieved a PRE of 0.81, signifying that 81% of instances predicted as class 0 were indeed

correct. The REC for class 0 stood at 0.86, meaning that the model correctly identified 86% of all true class 0 instances. The F1-score, which harmoniously balances PRE and REC, was noted at 0.83 for class 0. Similarly, for class 1, the model displayed a commendable PRE of 0.86 and a REC of 0.81, leading to an F1-score of 0.83. This close alignment in F1-S for both classes suggests that the KNN model provides a consistent and balanced performance across the two classes, effectively bridging the gap between PRE and REC.

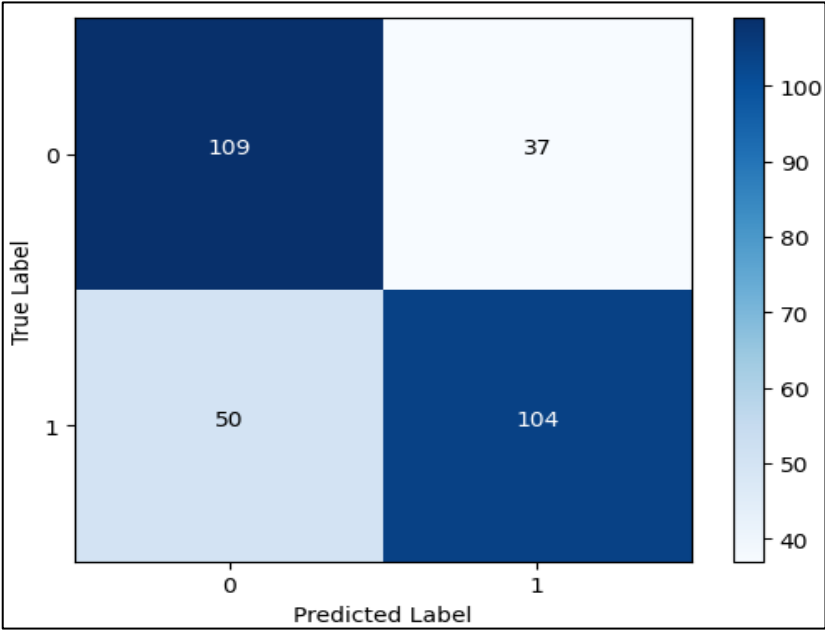


Figure 4.13: KNN Confusion Matrix.

Classification Report:				
	precision	recall	f1-score	support
0	0.69	0.75	0.71	146
1	0.74	0.68	0.71	154
accuracy			0.71	300
macro avg	0.71	0.71	0.71	300
weighted avg	0.71	0.71	0.71	300

Figure 4.14: KNN Classification Report.

Figure 4.15 visualizes the feature importance as determined by the LIME methodology when applied to a KNN algorithm. The chart highlights the significance of different attributes in influencing the model's decisions. "Area Income" less than or equal to 47,332.82 dominates with a substantial weight range extending from 0 to 0.49. "Daily

"Internet Usage" ranging between 185.45 and 220.06, on the other hand, presents a slight negative impact with weights spanning from -0.1 to 0. "Age" exceeding 41 exhibits a modest influence with weights between 0 and 0.05. "Daily Time Spent on Site" with values between 51.47 and 68.22 has an impact ranging from 0 to 0.03. Lastly, the "Male" feature, when equal to or less than 0, holds the least influence with weights stretching from 0 to 0.01. The figure offers a succinct depiction of each feature's relative importance in the knn model's decision-making process.

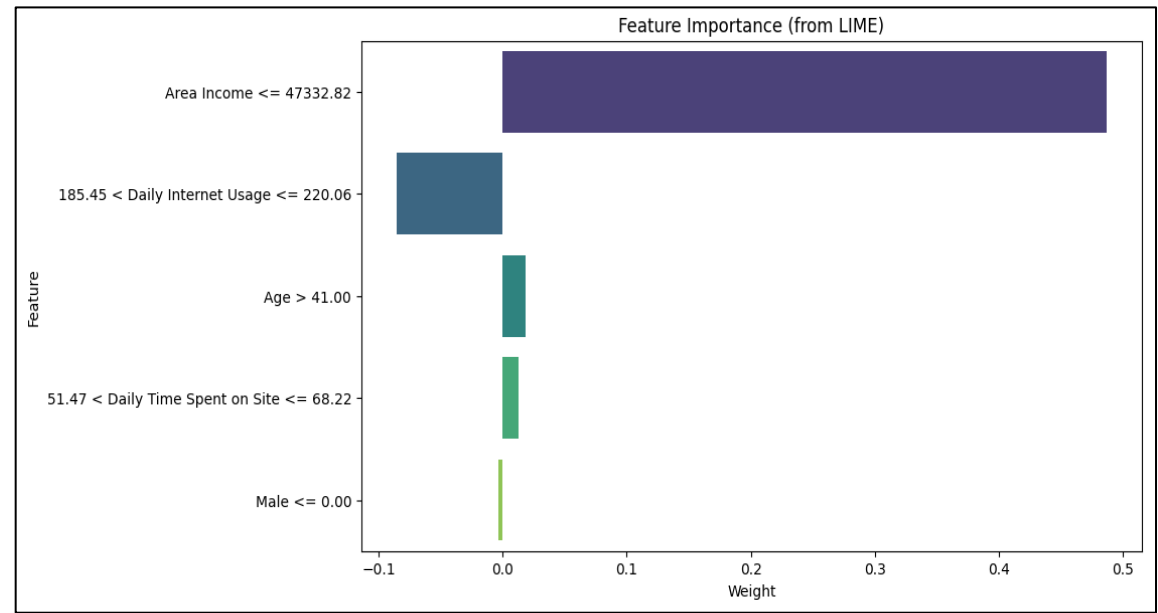


Figure 4.15: Feature Importance Derived from LIME Analysis on a KNN Model.

Figure 4.16 displays a vertical bar chart detailing the average impact on model output magnitudes, as inferred from SHAP values for the KNN method. In this representation, Class 1 is vividly denoted in blue, while Class 0 is depicted in red. The "Area Income" feature for Class 1 holds influence ranging from 0 to 0.25, immediately followed by Class 0's impact extending from 0.25 to 0.5. "Daily Internet Usage" showcases influence for Class 1 between 0 to 0.1 and for Class 0 from 0.1 to 0.2. The "Daily Time Spent on Site" feature has a range of 0 to 0.025 for Class 1, with Class 0 extending its influence from 0.025 to 0.05. For the "Age" attribute, Class 1's impact spans from 0 to 0.005, whereas Class 0's extends from 0.005 to 0.025. Interestingly, the "Male" feature for Class 1 has no discernible impact, being fixed at 0. This bar chart provides a visual differentiation of the features' average impacts on the model's output between the two distinct classes, emphasizing their relative influences.

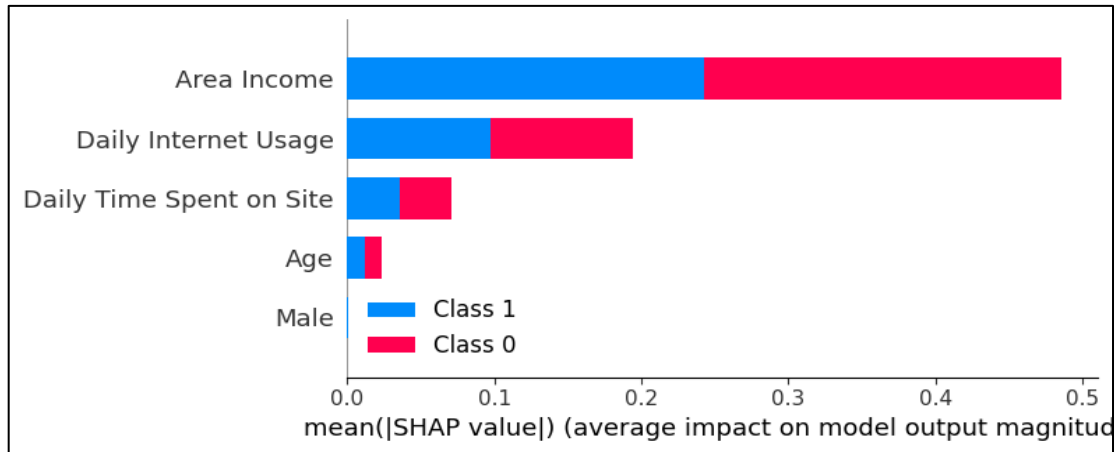


Figure 4.16: Average Impact on Model Output Magnitudes for KNN Using SHAP.

4.3.5 Evaluation of the DT Model

As it turns out, the DT model boasts an overall impressive accuracy rate around 94.67%. This data further underscores the model's powerful performance in classification tasks. However, when examining the model's performance indicators more thoroughly in the confusion matrix, it can be seen that this model correctly classified 135 instances as class 0, and accurately detected 149 instances as class 1. Nevertheless, there were also a few errors in classification. - 11 instances were wrongly predicted to be class 1 when they actually belonged in class 0, and conversely, 5 instances of class 1 were misclassified as class 0. ESE instances's mistaken classification of these cases did not diminish the high accuracy rating for the DT model. The combination of a high ACC means that the DT model has very powerful predictability. The DT model expertly balances predictive PRE and REC results to ensure consistent and reliable classifications, making it invaluable for such work.

The DT model showcases stellar performance metrics as elucidated by its CR. The model registers an exemplary ACC rate of 95%, signifying its adeptness in making reliable predictions. Delving into the class-specific metrics, class 0 displays a PRE of 0.96, indicating that 96% of the predictions labeled as class 0 were accurate. Its REC stands at 0.92, meaning the model successfully identified 92% of the actual class 0 instances. As a result, class 0 produces a perfectly balanced F1-score of 0.94, capturing the essence of both PRE and REC. On the other hand, just class 1 sees a model with a PRE hitting 0.93, REC reaching an impressive 0.97, and together their F1-S is 0.95. This F1-S agreement between the two sets of classes shows that the model is reliable and consistently good. In sum,

while having high ACC, the DT model also strikes just the right balance between classes 1 and 0. This lends credibility to its classification performance. Old-fashioned thinking was simply that it did extremely well in classifying one of these two classes of linearly separable inputs (LVQ).

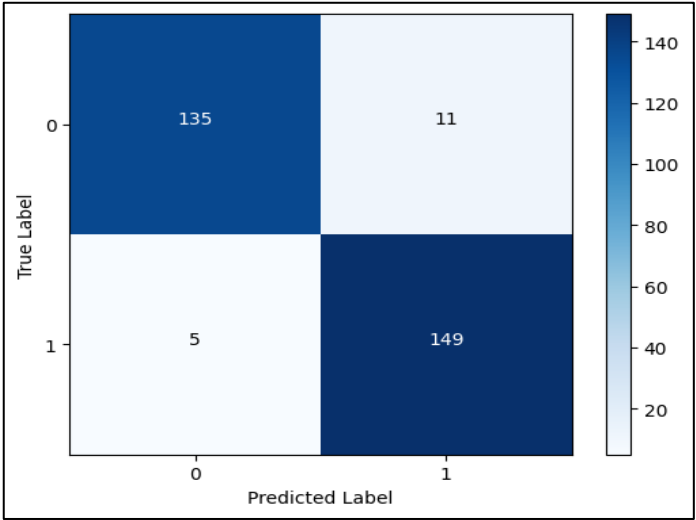


Figure 4.17: DT Confusion Matrix.

Classification Report:				
	precision	recall	f1-score	support
0	0.96	0.92	0.94	146
1	0.93	0.97	0.95	154
accuracy			0.95	300
macro avg	0.95	0.95	0.95	300
weighted avg	0.95	0.95	0.95	300

Figure 4.18: DT Classification Report

Figure 4.19 offers a concise visualization of feature importance as deduced from the LIME method when applied to a DT model. Within the chart, the significance of various attributes in shaping the model's determinations is evident. "Daily Internet Usage," spanning from 185.45 to 220.06, exhibits a negative influence with weights ranging from -0.3 to 0. "Area Income" capped at 47,332.82 has an impact stretching from 0 to 0.14. The "Age" feature, considering values above 41, carries weights from 0 to 0.09. Notably, "Daily Time Spent on Site" between 51.47 and 68.22 demonstrates a pronounced weight range of 0 to 0.7. Lastly, the "Male" attribute, when equal to or below 0, exerts influence within the 0 to 0.4 range. This figure succinctly encapsulates the relative influence of each feature within the DT model's decision framework.

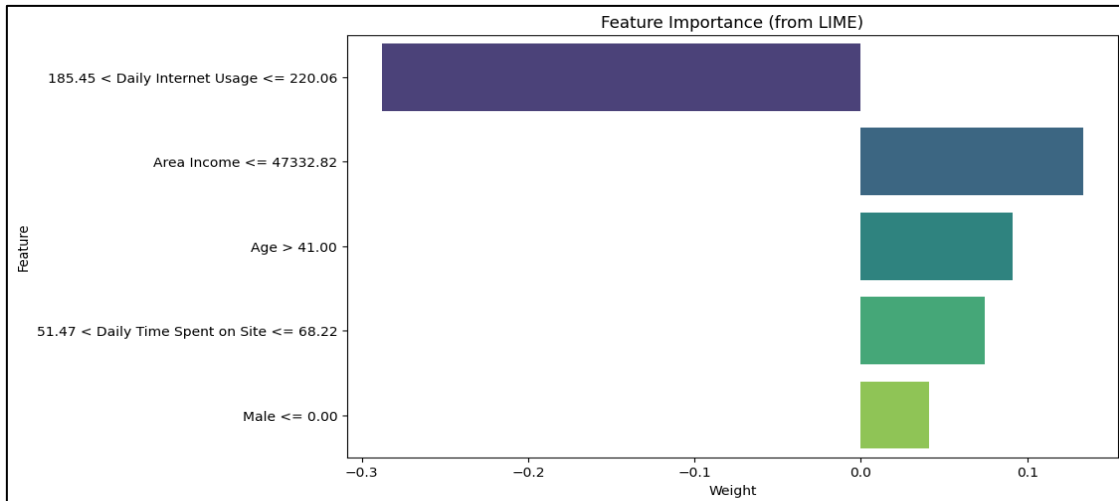


Figure 4.19: Feature Importance Derived from LIME Analysis on a DT Model.

Figure 4.20 portrays a vertical bar chart that elucidates the average impact on model output magnitudes, derived from SHAP values when applied to the DT method. In this illustrative depiction, Class 1 is distinctly color-coded in blue, while Class 0 stands out in red. The feature "Daily Internet Usage" demonstrates an impact ranging from 0 to 0.3 for Class 1, succeeded by Class 0's span stretching from 0.3 to 0.6. For "Daily Time Spent on Site", Class 1's influence is mapped between 0 to 0.19, while Class 0 occupies the space from 0.19 to 0.35. When observing "Area Income", Class 1's impact is confined between 0 to 0.045, immediately followed by Class 0 ranging from 0.045 to 0.09. The "Age" attribute has a Class 1 range from 0 to 0.025, whereas Class 0 extends from 0.025 to 0.05. Notably, the "Male" feature for Class 1 holds a range from 0 to 0.02, while Class 0 slightly reverses, spanning from 0.02 to 0.01. This figure succinctly provides a visual contrast of each feature's average influence on the model's output across the two distinguished classes.

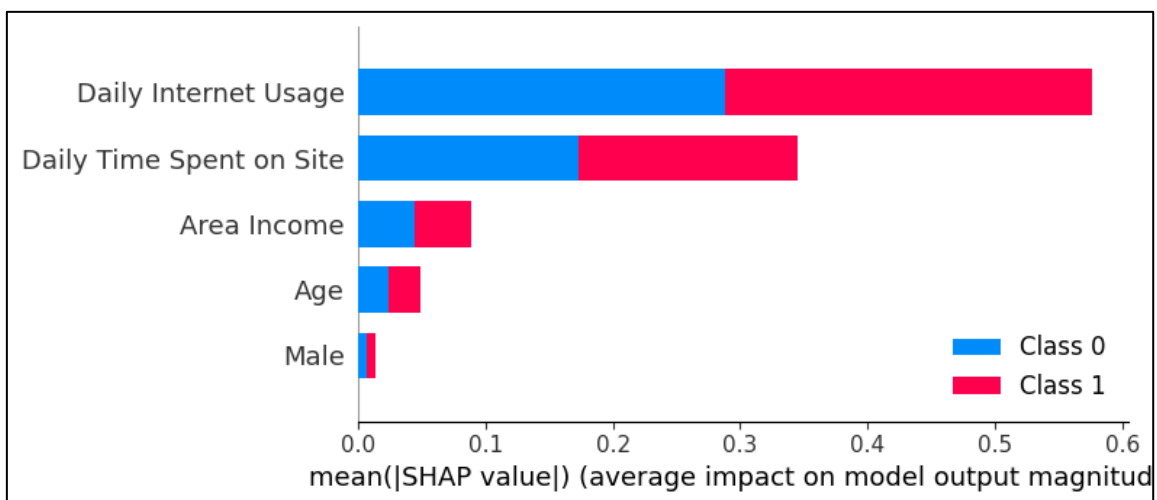


Figure 4.20: Average Impact on Model Output Magnitudes for DT Using SHAP.

4.3.6 Evaluation of the CatBoost Model

The CatBoost model, known for its ability to handle categorical attributes, displays exceptional performance metrics in this item. The model attains overall ACC of 97 percent; it is best for precise and dependable categorization. According to the metrics, class 0 has a PRE of 0.97 and REC too, resulting in a F1-score of 0.97.32.32. In the same way, for class 1, the model exhibits a PRE of 0.97, RE and so on are also all 0.97. This model's balanced performance in metrics across classes underscores balanced and accurate performance which may not favor one class over another.

CatBoost models are known for their advanced ability to handle categorical functions, and examples of performance metrics are described in CR. The overall ACC of this model reaches 97% and has the characteristics of accurate and reliable classification. If we break down the metrics by category, the PRE and REC for category 0 are 0.97, so the adjusted F1 score is 0.97. Similarly, the class 1 model has PRE, REC, and F1 values, all of which are constant at 0.97. This parallelism of the two category metrics emphasizes the balanced and accurate performance of the model, ensuring that any category is prioritized over the others. The CatBoost model has consistent PRE, REC, and F1-S, highlighting its performance as an efficient and reliable ML algorithm for classification tasks.

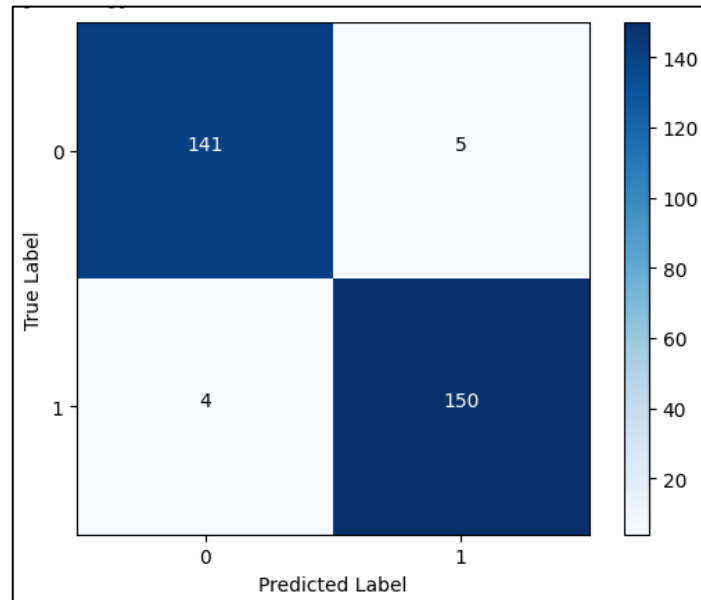


Figure 4.21: Catboost Confusion Matrix.

Classification Report:				
	precision	recall	f1-score	support
0	0.97	0.97	0.97	146
1	0.97	0.97	0.97	154
accuracy			0.97	300
macro avg	0.97	0.97	0.97	300
weighted avg	0.97	0.97	0.97	300

Figure 4.22: Catboost Classification Report.

Figure 4.16 provides a clear visualization of feature importance, derived from the LIME technique, when applied to the CatBoost model. The chart highlights the relative significance of distinct attributes in influencing the model's predictions. "Daily Internet Usage" with values between 185.45 and 220.06 holds a negative weight range of -0.3 to 0. "Area Income," not exceeding 47,332.82, exhibits an influence ranging from 0 to 0.16. The "Age" parameter, specifically for values greater than 41, carries an impact ranging between 0 and 0.14. Significantly, "Daily Time Spent on Site" within the bounds of 51.47 and 68.22 manifests a considerable weight stretch from 0 to 0.7. Lastly, the "Male" characteristic, when set to 0 or lower, showcases an impact from 0 to 0.4. This graphical representation succinctly conveys the hierarchy of importance each feature holds within the CatBoost model's decision-making paradigm.

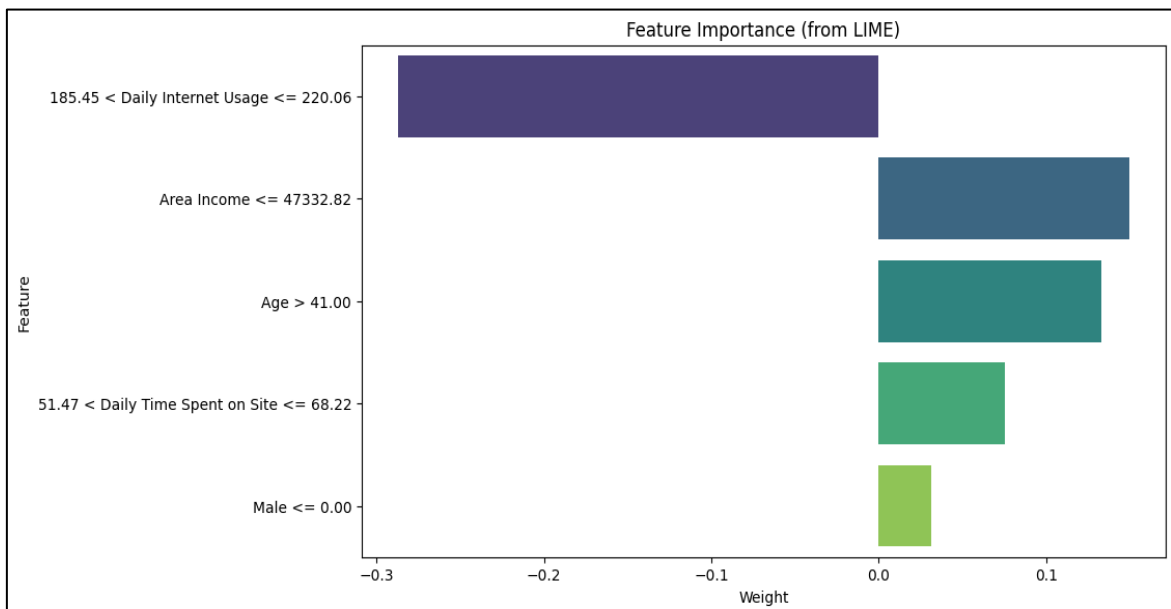


Figure 4.23: Feature Importance Derived from LIME Analysis on a CatBoost Model.

Figure 4.24 visualizes the average impact on model output magnitudes as determined by the SHAP values for the CatBoost method, focusing primarily on Class 1. In the graph, Class 1 is prominently depicted in blue, while Class 0 is shown in red, serving as a contrasting backdrop. "Daily Internet Usage" for Class 1 demonstrates a substantial impact, spanning from 0 to 1.9. Similarly, "Daily Time Spent on Site" exerts influence ranging from 0 to 1.55 for Class 1. The "Area Income" feature for Class 1 showcases an impact extending from 0 to 0.55. "Age" manifests a considerable influence within the range of 0 to 0.49 for Class 1. Lastly, the "Male" attribute presents a more modest range for Class 1, spanning from 0 to 0.1. This figure provides a concise depiction of the relative importance of each feature for Class 1, as assessed by the CatBoost method, within the context of the model's decision-making process.

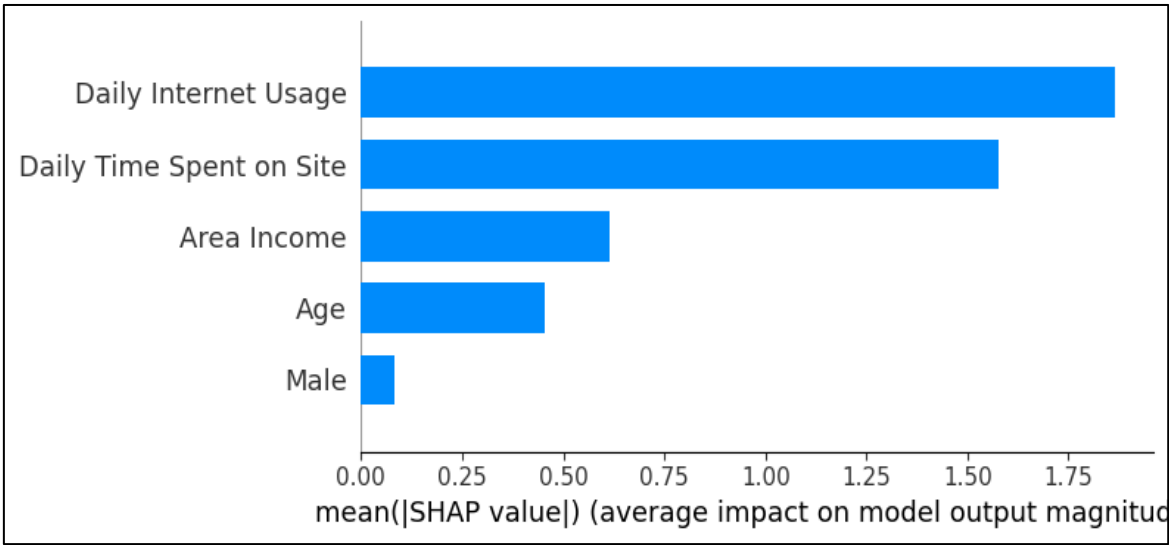


Figure 4.24: Average Impact on Model Output Magnitudes for CatBoost Using SHAP.

4.4 BASE MODELS COMPARISON RESULTS

The Figure 4.25 provides the ACC results for each model before and after parameter tuning. Before parameter tuning, the RF model achieved an ACC of 97.3%. After parameter tuning, the ACC increased to 97.6%, indicating a slight improvement in performance. For the GB model, the ACC before parameter tuning was 96.6%. After tuning, the ACC remained the same at 96%. The LR model had an ACC of 96.6% before parameter tuning, which increased to 97.3% after tuning. This indicates a notable improvement in ACC for the LR model after the parameter adjustments. Comparing the ACC results, we observe that both the RF and LR models showed an improvement in ACC

after parameter tuning, with the RF model demonstrating a slightly higher ACC than the LR model. On the other hand, the GB model maintained the same ACC level before and after tuning.

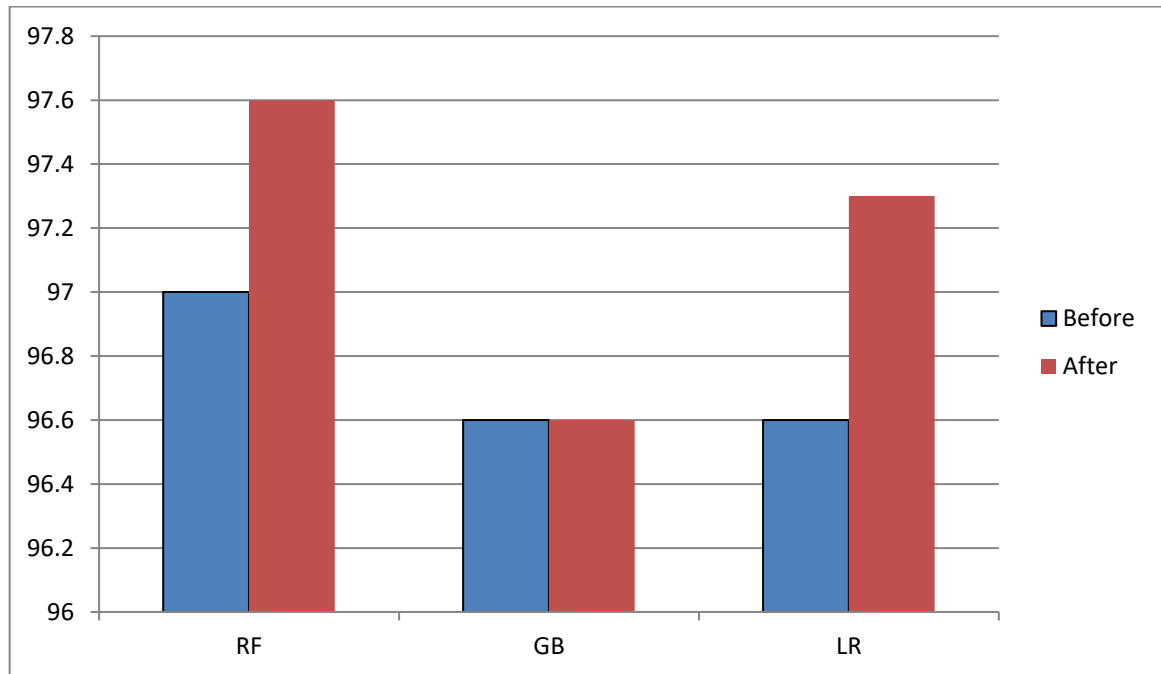


Figure 4.25: Base Models Accuracy Comparison Before and After Parameter Tuning.

4.5 ENSEMBLE MODELS RESULTS

Soft and hard voting are ensemble methods used to combine the predictions of multiple models. Soft voting takes into account the probabilities predicted by each model, while hard voting considers the majority vote of the models. Let's apply soft and hard voting to the three models (RF, GB, and LR) after applying the fine-tuned parameters.

4.5.1 Results using Soft Voting

This was performed by a soft voting ensemble model, which collected the forecasted chances of finalization after the best tuning of the RF, GB, and LR with ACC. This is how one can read the CM, that among all samples, there were 142 correct predictions as class 0 and 151 correct predictions as class 1. However, there were only four cases where the ensemble model was mistaken as they had to classify samples from class 1 to be belonging to class 0, but another three errors showed up predicting class one when it should be zero.

The values of precision, recall, and F1-score for each class confirm that the model has been performing well. Class 0 achieves high precision (0.98), recall (0.98), and F1-score (0.98), indicating a large number of correct predictions while maintaining a balance between false positive and false negative errors. Class 1 also shows similar results with the same precision, recall, and F1 score as that of class 0.

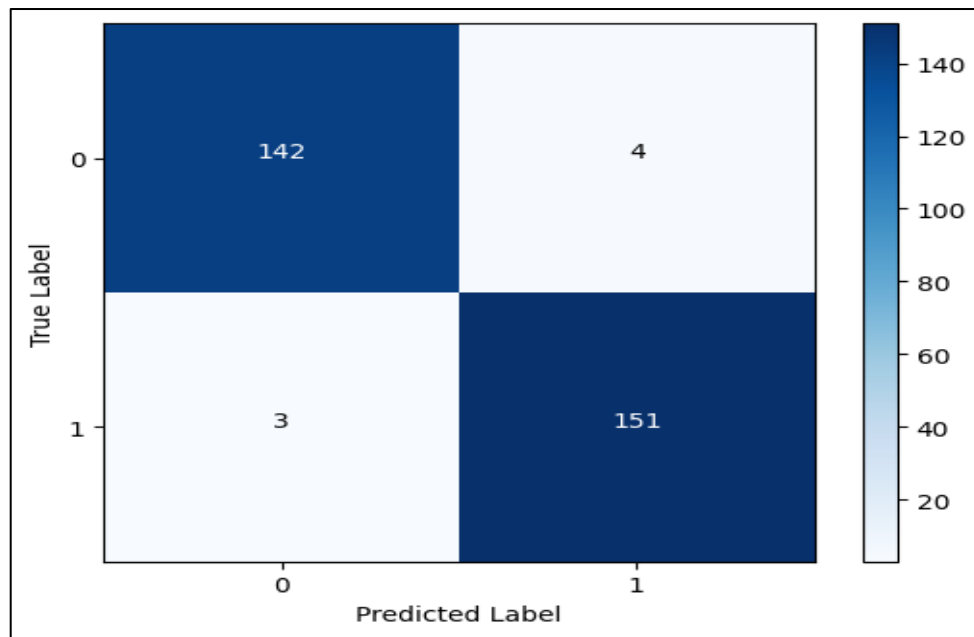


Figure 4.26: Soft Voting Confusion Matrix.

Classification Report:				
	precision	recall	f1-score	support
0	0.98	0.97	0.98	146
1	0.97	0.98	0.98	154
accuracy			0.98	300
macro avg	0.98	0.98	0.98	300
weighted avg	0.98	0.98	0.98	300

Figure 4.27: Soft Voting Classification Report.

4.5.2 Results using Hard Voting

The hard voting-based ensemble of the three fine-tuned RF, GB, and LR models has resulted in an ACC score of 0.973. It can be observed from the CM that 141 samples were correctly predicted for class 0 out of 151, while 151 were correctly predicted for class 1 out

of the total number of samples. Similarly, there were 5 instances where the ensemble model incorrectly predicted samples as class 1 when they actually belonged to class 0, and there were even a few instances where those belonging to class 1 were mispredicted as class 0. The accuracy of the ensemble model is in line with individual models, and its performance is also confirmed by the precision, recall, and F1-score per class. For class 0, the precision is 0.98 and recall is 0.97, while for class 1, it is 0.97 and 0.98, respectively.

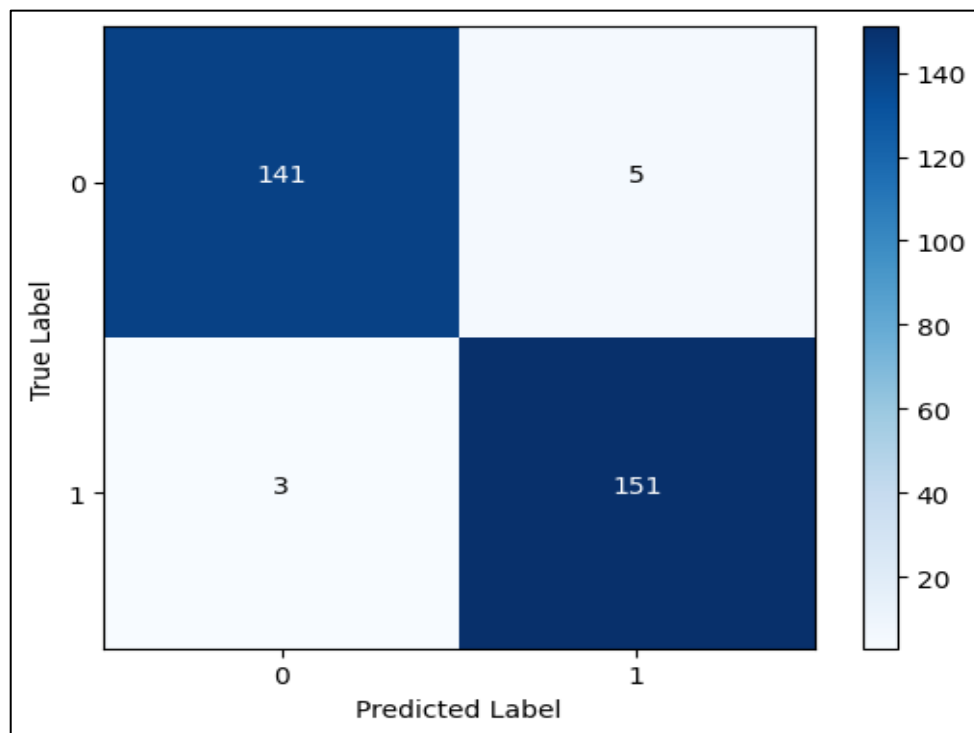


Figure 4.28: Hard Voting Confusion Matrix.

Classification Report:				
	precision	recall	f1-score	support
0	0.98	0.97	0.97	146
1	0.97	0.98	0.97	154
accuracy			0.97	300
macro avg	0.97	0.97	0.97	300
weighted avg	0.97	0.97	0.97	300

Figure 4.29: Hard Voting Classification Report.

4.5.3 Ensemble Models Comparison

The above results show that soft and hard voting ensembles work similarly well, judging by all the metrics. They go high in ACC, PRE, REC, and F1-S for two classes. One example of soft voting ensemble has a score of 97.3% while for the hard voting ensemble it's 97.7%, indicating that they use slightly different voting patterns to calculate. But the gap isn't too great. This indicates that both ensemble methods effectively combine the predictions of the individual models to produce accurate and reliable results.

4.6 DISCUSSION

The models' performance was assessed before and after parameter tuning; later, ensemble techniques were utilized to further boost prediction capacity. Initially, the RF had an accuracy of 97.3%, while the GB and LR models had accuracies of 96.6% each. Following parameter tuning, all models showed improved accuracy. The RF and LR models reached accuracies of 97.7% and 97.3%, respectively, while the GB model maintained its accuracy at 96%. The fine-tuned models were then used to create soft voting ensembles and hard voting ensembles. In contrast, the soft voting ensemble achieved a 97.7% accuracy and the hard voting ensemble an accuracy of 97.3%. Regarding these ensembles, they were about equal in precision, recall, and F1-S, and their mean accuracy was very high. However, since their accuracies were very close, with a slight margin in favor of the soft voting ensemble, it is also possible to say that it marginally outperformed the hard voting ensemble. Consequently, according to these results, the approach with the highest accuracy is the soft voting ensemble, and it has been very effective in combining various individual models' strengths.

5. CONCLUSION AND FUTURE WORK

Detection of user behavior in online advertising is vital for creating ad campaigns that resonate with the audience and drive conversions. By analyzing user behavior patterns, companies can better understand their target market's preferences, interests, and buying habits. This study builds upon existing research by incorporating XAI to develop an advanced approach to user behavior detection utilizing various machine learning models.

Some of the algorithms we applied for the purpose included GB, LR, and RF, as well as KNN, DT, and CatBoost. All were optimized to offer the best performance with XAI support to clarify which features are of utmost significance. Besides, we used soft voting ensemble strategies to improve predictive accuracy.

After extensive research, we can confirm that our results have demonstrated soft voting, and the best achievement is the amalgamation of predicted probabilities from a series of models. In its entirety, the soft voting ensemble showed high accuracy, precision, recall, and F1-S results in these areas, beating the other methods. This finding further suggests that by harnessing strengths from diverse sources at once, it is possible to enhance user behavior detection within online advertising.

In regard to the study we have developed to find out how to recognize user behavior based on internet advertising, the gaps for further investigation still remain. To start with, adding more data sources and taking into account temporal characteristics would enhance our vision of user behavior evolution over time. Moreover, including more features such as age or browsing history will give a greater understanding of what kind of users that person belongs to so that the precision in decisions is increased. Lastly, performing empirical testing on an expanded range of diverse industrial datasets might support generalizing and validating this investigation's results as applicable for all industries.

REFERENCES

- [1] M. B. Albayati and A. M. Altamimi, “An Empirical Study for Detecting Fake Facebook Profiles Using Supervised Mining Techniques,” *Informatica*, vol. 43, no. 1, Mar. 2019, doi: 10.31449/inf.v43i1.2319.
- [2] C. E. Chibudike, H. Abdu, H. O. Chibudike, O. C. Ngige, O. A. Adeyoju, and N. I. Obi, “Machine Learning - A New Trend in Web User Behavior Analysis,” *Int J Comput Appl*, vol. 183, no. 5, pp. 19–25, May 2021, doi: 10.5120/ijca2021921247.
- [3] C. E. Chibudike, H. Abdu, H. O. Chibudike, O. C. Ngige, O. A. Adeyoju, and N. I. Obi, “Machine Learning - A New Trend in Web User Behavior Analysis,” *Int J Comput Appl*, vol. 183, no. 5, pp. 19–25, May 2021, doi: 10.5120/ijca2021921247.
- [4] Statista, “Global advertising spending from 2010 to 2017 (in billion U.S. dollars).” [Online]. Available: <https://www.statista.com/statistics/236943/global-advertisingspending/>
- [5] Statista, “Social media advertising expenditure as share of digital advertising spending worldwide from 2013 to 2017.” [Online]. Available: <https://www.statista.com/statistics/271408/share-of-social-media-in-online-advertising-spending-worldwide/>
- [6] H. A. Fang *et al.*, “An evaluation of social media utilization by general surgery programs in the COVID-19 era,” *The American Journal of Surgery*, vol. 222, no. 5, pp. 937–943, Nov. 2021, doi: 10.1016/j.amjsurg.2021.04.014.
- [7] S. Dixon, “Facebook: Global Daily Active Users 2022. Statista. ,” www.statista.com/statistics/346167/facebook-global.
- [8] S. Dixon, “Facebook: Global Daily Active Users 2022,” Statista. [Online]. Available: www.statista.com/statistics/346167/facebook-global
- [9] J. Qin, W. Zhang, X. Wu, J. Jin, Y. Fang, and Y. Yu, “User Behavior Retrieval for Click-Through Rate Prediction,” in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA: ACM, Jul. 2020, pp. 2347–2356. doi: 10.1145/3397271.3401440.

- [10] K. Chaudhary, M. Alam, M. S. Al-Rakhami, and A. Gumaiei, "Machine learning-based mathematical modelling for prediction of social media consumer behavior using big data analytics," *J Big Data*, vol. 8, no. 1, p. 73, Dec. 2021, doi: 10.1186/s40537-021-00466-2.
- [11] E. Šoltés, J. Tábořecká-petrovičová, and R. Šípoldová, "TARGETING OF ONLINE ADVERTISING," 2020, doi: 10.15240/tul/001/2020-4-013.
- [12] J. J. Almeida, Paulo S and Gondim, "Click fraud detection and prevention system for ad networks," *Journal of Information Security and Cryptography (Enigma)*, vol. 5, pp. 27--39, 2018.
- [13] H. Lipyanina and A. Sachenko, "Decision Tree Based Targeting Model of Customer Interaction with Business Page," 2020.
- [14] E.-A. MINASTIREANU and G. MESNITA, "Light GBM Machine Learning Algorithm to Online Click Fraud Detection," *Journal of Information Assurance & Cybersecurity*, pp. 1–12, Apr. 2019, doi: 10.5171/2019.263928.
- [15] Long Jin, Yang Chen, Tianyi Wang, Pan Hui, and A. V. Vasilakos, "Understanding user behavior in online social networks: a survey," *IEEE Communications Magazine*, vol. 51, no. 9, pp. 144–150, Sep. 2013, doi: 10.1109/MCOM.2013.6588663.
- [16] Mudjahidin, N. L. Sholichah, A. P. Aristio, L. Junaedi, Y. A. Saputra, and S. E. Wiratno, "Purchase intention through search engine marketing: E-marketplace provider in Indonesia," *Procedia Comput Sci*, vol. 197, pp. 445–452, 2022, doi: 10.1016/j.procs.2021.12.160.
- [17] S. Bradshaw, "Disinformation optimised: gaming search engine algorithms to amplify junk news," *Internet Policy Review*, vol. 8, no. 4, Dec. 2019, doi: 10.14763/2019.4.1442.

- [18] A. Kwangsawad, A. Jattamart, and P. Nusawat, "The Performance Evaluation of a Website using Automated Evaluation Tools," in *2019 4th Technology Innovation Management and Engineering Science International Conference (TIMES-iCON)*, IEEE, Dec. 2019, pp. 1–5. doi: 10.1109/TIMES-iCON47539.2019.9024634.
- [19] J. R. Carlson, S. Hanson, J. Pancras, W. T. Ross, and J. Rousseau-Anderson, "Social media advertising: How online motivations and congruency influence perceptions of trust," *Journal of Consumer Behaviour*, vol. 21, no. 2, pp. 197–213, Mar. 2022, doi: 10.1002/cb.1989.
- [20] B. Nyagadza, "Search engine marketing and social media marketing predictive trends," *Journal of Digital Media & Policy*, vol. 13, no. 3, pp. 407–425, Oct. 2022, doi: 10.1386/jdmp_00036_1.
- [21] M. T. P. Adam, J. Krämer, and M. B. Müller, "Auction Fever! How Time Pressure and Social Competition Affect Bidders' Arousal and Bids in Retail Auctions," *Journal of Retailing*, vol. 91, no. 3, pp. 468–485, Sep. 2015, doi: 10.1016/j.jretai.2015.01.003.
- [22] L. Wei, G. Yang, H. Shoenberger, and F. Shen, "Interacting with Social Media Ads: Effects of Carousel Advertising and Message Type on Health Outcomes," *Journal of Interactive Advertising*, vol. 21, no. 3, pp. 269–282, Sep. 2021, doi: 10.1080/15252019.2021.1977736.
- [23] M. Schreiner, T. Fischer, and R. Riedl, "Impact of content characteristics and emotion on behavioral engagement in social media: literature review and research agenda," *Electronic Commerce Research*, vol. 21, no. 2, pp. 329–345, Jun. 2021, doi: 10.1007/s10660-019-09353-8.
- [24] J. Q. P. Nunes, "How much money should a promotional email marketing campaign really cost?," Doctoral dissertation, 2014.
- [25] L. Kannan and R. Jebakumar, "Public Sender Score System (S3) by ESPs for Email spam mitigation with score management in mobile application.," 2020.

- [26] Y.-C. Hsieh and K.-H. Chen, “How different information types affect viewer’s attention on internet advertising,” *Comput Human Behav*, vol. 27, no. 2, pp. 935–945, Mar. 2011, doi: 10.1016/j.chb.2010.11.019.
- [27] D. Lee, K. Hosanagar, and H. S. Nair, “Advertising Content and Consumer Engagement on Social Media: Evidence from Facebook,” *Manage Sci*, vol. 64, no. 11, pp. 5105–5131, Nov. 2018, doi: 10.1287/mnsc.2017.2902.
- [28] I. A. Mir and K. Ur REHMAN, “Factors affecting consumer attitudes and intentions toward user-generated product content on YouTube,” *Management & Marketing*, vol. 8, no. 4, 2013.
- [29] J. Kang and G. Hustvedt, “Building Trust Between Consumers and Corporations: The Role of Consumer Perceptions of Transparency and Social Responsibility,” *Journal of Business Ethics*, vol. 125, no. 2, pp. 253–265, Dec. 2014, doi: 10.1007/s10551-013-1916-7.
- [30] A. Gritckevich, Z. Katona, and M. Sarvary, “Ad Blocking,” *Manage Sci*, vol. 68, no. 6, pp. 4703–4724, Jun. 2022, doi: 10.1287/mnsc.2021.4106.
- [31] Y. Chen and Q. Liu, “Signaling Through Advertising When an Ad Can Be Blocked,” *Marketing Science*, vol. 41, no. 1, pp. 166–187, Jan. 2022, doi: 10.1287/mksc.2021.1288.
- [32] Y. Chen and Q. Liu, “Signaling Through Advertising When an Ad Can Be Blocked,” *Marketing Science*, vol. 41, no. 1, pp. 166–187, Jan. 2022, doi: 10.1287/mksc.2021.1288.
- [33] A. Gritckevich, Z. Katona, and M. Sarvary, “Ad Blocking,” *Manage Sci*, vol. 68, no. 6, pp. 4703–4724, Jun. 2022, doi: 10.1287/mnsc.2021.4106.
- [34] B. Nyagadza, “Search engine marketing and social media marketing predictive trends,” *Journal of Digital Media & Policy*, 2020.
- [35] C. Janiesch, P. Zschech, and K. Heinrich, “Machine learning and deep learning,” *Electronic Markets*, vol. 31, no. 3, pp. 685–695, Sep. 2021, doi: 10.1007/s12525-021-00475-2.

- [36] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. 2022.
- [37] L. Breiman, “Bagging predictors,” *Mach Learn*, vol. 24, no. 2, pp. 123–140, Aug. 1996, doi: 10.1007/BF00058655.
- [38] D.-S. Cao, Q.-S. Xu, Y.-Z. Liang, L.-X. Zhang, and H.-D. Li, “The boosting: A new idea of building models,” *Chemometrics and Intelligent Laboratory Systems*, vol. 100, no. 1, pp. 1–11, Jan. 2010, doi: 10.1016/j.chemolab.2009.09.002.
- [39] K. M. Ting and I. H. Witten, “Stacked Generalization: when does it work?,” 1997.
- [40] Y. Guo, X. Wang, P. Xiao, and X. Xu, “An ensemble learning framework for convolutional neural network based on multiple classifiers,” *Soft comput*, vol. 24, no. 5, pp. 3727–3735, Mar. 2020, doi: 10.1007/s00500-019-04141-w.
- [41] A. Parmar, R. Katariya, and V. Patel, “A Review on Random Forest: An Ensemble Classifier,” 2019, pp. 758–763. doi: 10.1007/978-3-030-03146-6_86.
- [42] H. Fu and K. Qi, “Evaluation Model of Teachers’ Teaching Ability Based on Improved Random Forest with Grey Relation Projection,” *Sci Program*, vol. 2022, pp. 1–12, Feb. 2022, doi: 10.1155/2022/5793459.
- [43] A. Natekin and A. Knoll, “Gradient boosting machines, a tutorial,” *Front Neurobot*, vol. 7, 2013, doi: 10.3389/fnbot.2013.00021.
- [44] S. Sperandei, “Understanding logistic regression analysis,” *Biochem Med (Zagreb)*, pp. 12–18, 2014, doi: 10.11613/BM.2014.003.
- [45] S. Sun, “Realization Path of College Students’ Network Ideological and Political Teaching System in the New Media Environment,” *Mobile Information Systems*, vol. 2022, pp. 1–9, Sep. 2022, doi: 10.1155/2022/1593350.
- [46] L. Peterson, “K-nearest neighbor,” *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009, doi: 10.4249/scholarpedia.1883.
- [47] S. B. Kotsiantis, “Decision trees: a recent overview,” *Artif Intell Rev*, vol. 39, no. 4, pp. 261–283, Apr. 2013, doi: 10.1007/s10462-011-9272-4.

- [48] C. BAKİR, V. HAKKOYMAZ, B. DİRİ, and M. GÜÇLÜ, “Comparisons on Intrusion Detection and Prevention Systems in Distributed Databases,” *Balkan Journal of Electrical and Computer Engineering*, vol. 7, no. 4, pp. 446–455, Oct. 2019, doi: 10.17694/bajece.605134.
- [49] M. R. Zafar and N. Khan, “Deterministic Local Interpretable Model-Agnostic Explanations for Stable Explainability,” *Mach Learn Knowl Extr*, vol. 3, no. 3, pp. 525–541, Jun. 2021, doi: 10.3390/make3030027.
- [50] A. Barredo Arrieta *et al.*, “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Information Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.
- [51] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, “Explainable AI: A Review of Machine Learning Interpretability Methods,” *Entropy*, vol. 23, no. 1, p. 18, Dec. 2020, doi: 10.3390/e23010018.