Unlocking Customer Behavior for Effective Marketing

Final project report

Team Name: Data Enthusiasts

Team Members:(leader)Amanullah and (teammate)Daniyal Hyder

Assigned by: Miss Reema Memon

Dataset Name: Digital Marketing Dataset

Colab Notebook

GitHub Repository

Report information:

This report is of our final project PITP course, focuses on understanding customer behavior and optimizing marketing strategies by analyzing engagement data, loyalty points, and campaign metrics. It addresses key questions like identifying customers at risk of leaving, segmenting customers based on engagement behavior to develop targeted marketing strategies, and evaluating the effectiveness of Social Media and PPC campaigns through hypothesis testing. The report also assesses the overall effectiveness of various marketing channels and campaigns. Finally, it explores the use of classification models to predict customer conversion based on demographic, campaign performance, and engagement data, offering insights to enhance retention strategies and marketing efficiency.

Background of Project:

In a digital-first world where both e-commerce and physical stores depend on digital marketing to connect with customers, digital marketers face complex challenges in making data-driven decisions. This project identifies key areas of confusion and difficulty commonly experienced in digital marketing and aims to provide solutions through targeted project goals that may help marketers navigate today's competitive landscape.

Inspiration:

The inspiration behind this project comes from a desire to tackle a challenging, impactful problem that could genuinely benefit digital marketers. We wanted a project that would not only apply the full range of skills we've acquired throughout this course but also put our knowledge into action, and create something valuable and relevant for the industry.

Problem Statement:

The goal of this Project is to understand customer behavior and evaluate the effectiveness of various marketing strategies. Specifically, we aim to predict customer conversion based on various factors such as demographics, campaign performance, and engagement metrics. To address this, we approached the problem through five (5) key questions mentioned below with their answers, applying different techniques and models to answer them.

About Dataset:

Context:

As digital marketing becomes increasingly data-driven, having accurate and comprehensive data is essential for analyzing customer behavior and optimizing marketing strategies. While some datasets focus on individual marketing channels or user interactions, comprehensive datasets that integrate engagement, demographic, and campaign metrics are relatively rare. This project's dataset bridges that gap, offering an extensive look at customer engagement, campaign performance, and conversion data.

To ensure the highest quality, we filtered this data on Kaggle using criteria like usability and completeness, identifying it as one of the most accurate and well-structured options available.

Content:

The dataset contains **2000** rows and **20** columns, detailed information across multiple aspects of digital marketing and customer engagement, making it suitable for analyzing customer behavior and the effectiveness of marketing strategies. Here is a breakdown of the features included:

Demographic Information:

- CustomerID: Unique identifier for each customer.
- Age: Age of the customer.
- Gender: Gender of the customer (Male/Female).
- Income: The customer's annual income in USD.

Marketing-Specific Variables:

- CampaignChannel: The channel through which the marketing campaign is delivered (options include Email, Social Media, SEO, PPC, Referral).
- CampaignType: Type of the marketing campaign (Awareness, Consideration, Conversion, Retention).
- AdSpend: The amount spent on the marketing campaign in USD.
- ClickThroughRate: Rate at which customers click on the marketing content.
- ConversionRate: Rate at which clicks convert to desired actions (e.g., purchases).
- AdvertisingPlatform: Platform through which advertisements were displayed (Confidential).
- AdvertisingTool: Tool used for managing and tracking advertisements (Confidential).

Customer Engagement Variables:

- WebsiteVisits: Number of visits to the website.
- PagesPerVisit: Average number of pages visited per session.
- TimeOnSite: Average time spent on the website per visit (in minutes).
- SocialShares: Number of times the marketing content was shared on social media.
- EmailOpens: Number of times marketing emails were opened.
- EmailClicks: Number of times links in marketing emails were clicked.

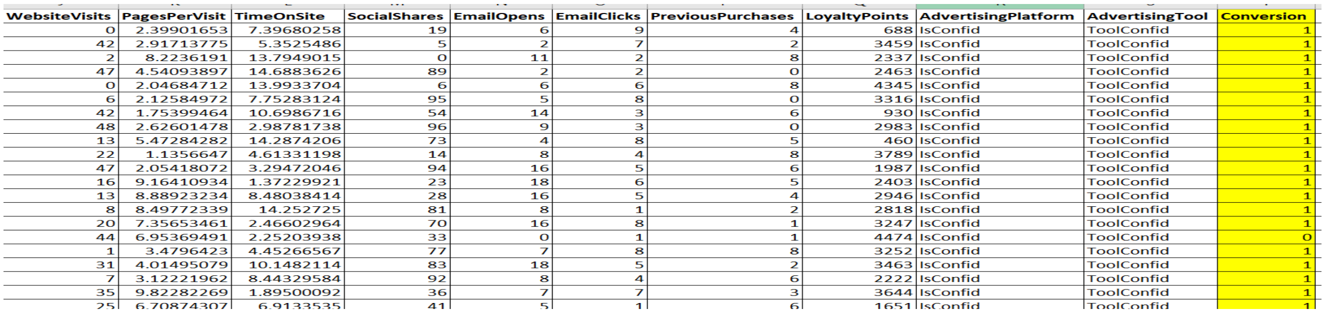
Historical Data

- PreviousPurchases: Number of previous purchases made by the customer.
- LoyaltyPoints: Number of loyalty points accumulated by the customer.

Target Variable:

- Conversion: Binary variable indicating whether the customer converted (1) or not (0). This is the primary outcome we aim to predict and analyze, representing successful customer actions influenced by marketing efforts

Target variable



WebsiteVisits	PagesPerVisit	TimeOnSite	SocialShares	EmailOpens	EmailClicks	PreviousPurchases	LoyaltyPoints	AdvertisingPlatform	AdvertisingTool	Conversion
0	2.39901653	7.39680258	19	6	9	4	688	IsConfid	ToolConfid	1
42	2.91713775	5.3525486	5	2	7	2	3459	IsConfid	ToolConfid	1
2	8.2236191	13.7949015	0	11	2	8	2337	IsConfid	ToolConfid	1
47	4.54093897	14.6883626	89	2	2	0	2463	IsConfid	ToolConfid	1
0	2.04684712	13.9933704	6	6	6	8	4345	IsConfid	ToolConfid	1
6	2.12584972	7.75283124	95	5	8	0	3316	IsConfid	ToolConfid	1
42	1.75399464	10.6986716	54	14	3	6	930	IsConfid	ToolConfid	1
48	2.62601478	2.98781738	96	9	3	0	2983	IsConfid	ToolConfid	1
13	5.47284282	14.2874206	73	4	8	5	460	IsConfid	ToolConfid	1
22	1.1356647	4.61331198	14	8	4	8	3789	IsConfid	ToolConfid	1
47	2.05418072	3.29472046	94	16	5	6	1987	IsConfid	ToolConfid	1
16	9.16410934	1.37229921	23	18	6	5	2403	IsConfid	ToolConfid	1
13	8.88923234	8.48038414	28	16	5	4	2946	IsConfid	ToolConfid	1
8	8.49772339	14.252725	81	8	1	2	2818	IsConfid	ToolConfid	1
20	7.35653461	2.46602964	70	16	8	1	3247	IsConfid	ToolConfid	1
44	6.95369491	2.25203938	33	0	1	1	4474	IsConfid	ToolConfid	0
1	3.4796423	4.45266567	77	7	8	8	3252	IsConfid	ToolConfid	1
31	4.01495079	10.1482114	83	18	5	2	3463	IsConfid	ToolConfid	1
7	3.12221962	8.44329584	92	8	4	6	2222	IsConfid	ToolConfid	1
35	9.82282269	1.89500092	36	7	7	3	3644	IsConfid	ToolConfid	1
25	6.70874307	6.9133535	41	5	1	6	1651	IsConfid	ToolConfid	1

Data Sourced:

We sourced this data from Kaggle, one of the most renowned and reliable platforms for public datasets.

Step 1: Overview of data columns and their types

1. **Categorical Columns (Qualitative Information)**

Examples: Gender, CampaignChannel, CampaignType

Usefulness:

- Categorical data can provide grouping or segmentation (e.g., analyzing conversion rates based on CampaignChannel).
- These columns help create classes or labels that influence outcomes, such as identifying which campaign type yields better conversions.

2. **Numerical Columns (Quantitative Information)**

   Examples: Age, Income, WebsiteVisits, PagesPerVisit, TimeOnSite, EmailOpens, EmailClicks, PreviousPurchases, LoyaltyPoints.

   Usefulness:

- ○ Numerical data provides measurable characteristics that can be analyzed for trends (e.g., how website visits relate to conversion rates).
- ○ Important for regression tasks where continuous values (like AdSpend or ConversionRate) are predicted.

3. Useless columns for the project

customer_id: This is likely just an identifier and does not provide predictive information regarding conversion.

advertising_platform: If it only contains a single unique value ('IsConfid'), it does not provide any useful information for prediction and should be dropped.

advertising_tool: Similar to the above, if it contains only one unique value ('ToolConfid'), it does not contribute to the prediction

Step 2: Libraries with Version Used

- **Pandas** (version: 2.2.2): For data manipulation and analysis, including data cleaning and preprocessing.
- **NumPy** (version: 1.26.4): For handling numerical operations and managing arrays.
- **Matplotlib** (version: 3.8.0): To create static plots for visualizing data distributions and trends.
- **Seaborn** (version: 0.13.2): For statistical data visualization and creating detailed plots.
- **Plotly** (version: 5.24.1): To generate interactive visualizations for dynamic exploration of data.

- **Scikit-Learn** (version: 1.5.2): A comprehensive library of machine learning algorithms and tools for model building and evaluation.
- **XGBoost** (version: xgboost): For gradient boosting algorithms, adding enhanced model performance with tree-based methods.
- **Statsmodels** (version: 0.14.4): To perform statistical modeling and hypothesis testing.
- **Scipy** (version: 1.13.1): For scientific computing and supporting statistical operations.
- **TensorFlow** (version: 2.17.0): A deep learning library used to explore complex patterns and relationships in the data.

Step 3: Data Preparation and Pre-processing

Data Wrangling and Cleaning:

Standardized and normalized: the data for consistency across all columns. This included:

- Converting all text data to lowercase for case normalization.
- Replacing spaces with underscores in column headers to maintain a consistent format.
- Removing unnecessary symbols or punctuation from text fields.

Handling Missing Values:

- Checked for null values using the **isnull().sum()** function. No missing values were found in the dataset.

Dropping Unnecessary Columns:

Removed columns that did not contribute predictive value:

- **customer_id**: Used only as an identifier, irrelevant for prediction.
- **advertising_platform** and **advertising_tool**: Contained a single unique value, offering no variability for prediction.
- Checking for Duplicates:
- Identified and confirmed that there were no duplicate rows in the dataset, ensuring uniqueness.

Outliers:

- Analyzed numeric data for outliers; no significant outliers were detected, and the distribution of values was approximately normal.

Feature Engineering / Transformation

Combining Features:

- Total Page Views: Combined **website_visits** and **pages_per_visit** into a single feature, **total_page_views**, to represent overall user engagement. This was done by multiplying the two features, providing a clearer indicator of engagement intensity, which may better correlate with conversion likelihood.
- Email Engagement: Combined **email_opens** and **email_clicks** into a single feature, **email_engagement**, to provide a unified view of email interaction. This was achieved by renaming **email_opens** to **email_engagement** and dropping **email_clicks**, making the variable more representative of the overall email engagement behavior.

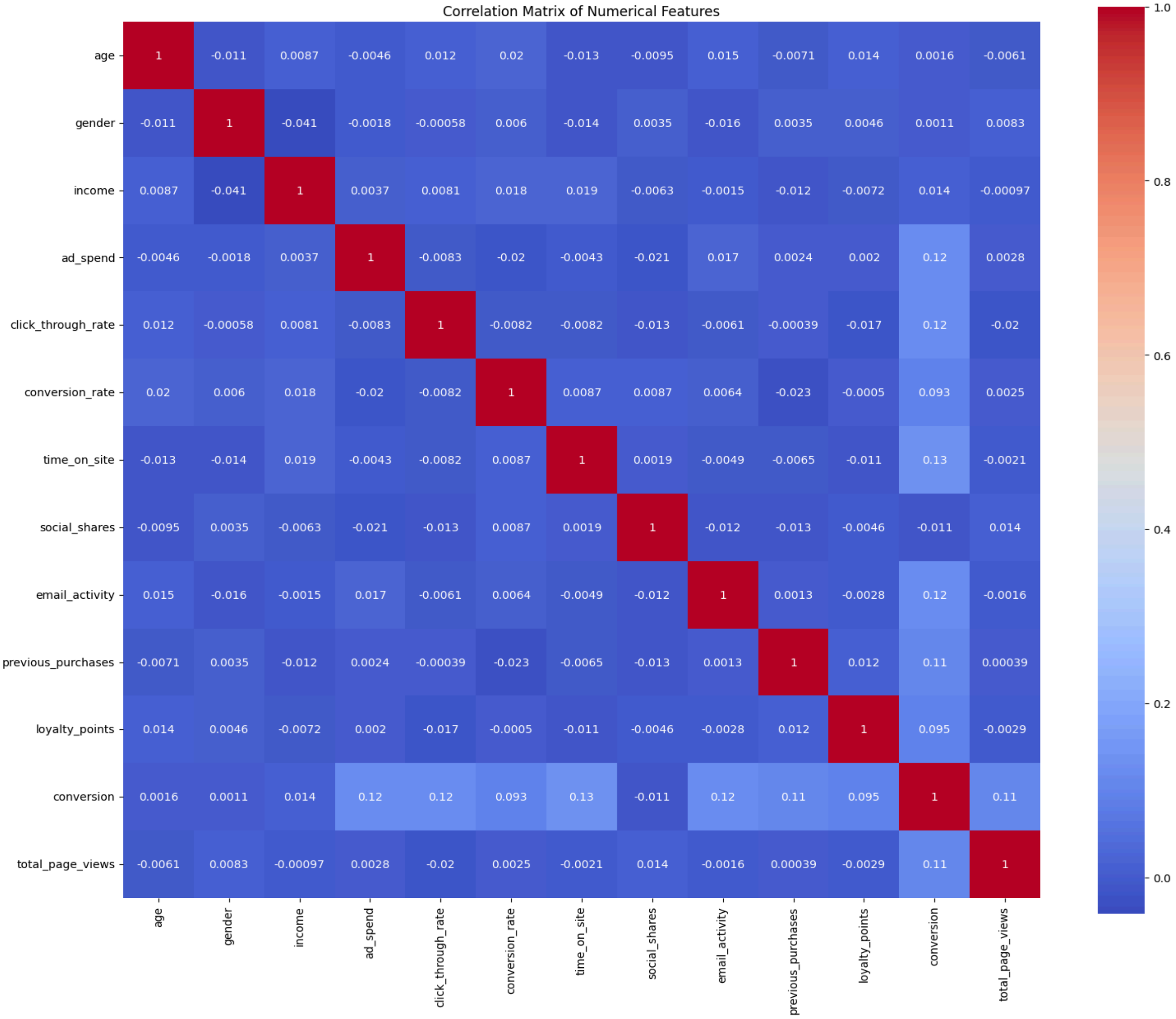
Feature Encoding:

- The **gender** column was encoded into binary values, where **0** represents "Male" and 1 represents "Female", allowing the model to handle this categorical variable effectively.

Step 4: Explore Relationships Between Columns

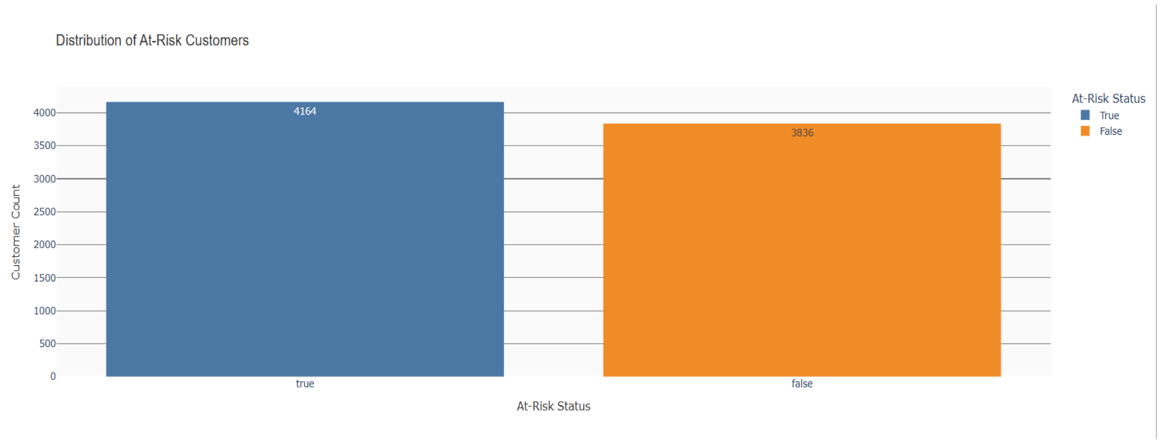
- **Correlation Analysis:**
  The correlation matrix reveals that there are generally weak correlations between the numerical features in the dataset, indicating that there are no strong linear relationships among most variables. For example, while conversion rate shows a notable correlation with conversion (~0.93), other features such as ad spend, income, click-through rate, and previous purchases exhibit weak correlations with each other and with conversion outcomes. This lack of strong linear associations suggests that the data is likely non-linear, meaning that more complex models may be needed to capture the underlying patterns effectively.

- The columns in the graph further demonstrate this **nonlinear** nature, as the majority of values in the heatmap hover around zero, showing minimal or no linear correlation. As a result, traditional linear models may not fully capture the relationships in this dataset, making techniques like decision trees, random forests, neural networks, Gradient Boosting, XGBoost, or LightGBM more suitable for accurately predicting customer behavior and conversions.



Problem1: Which customers are at risk of leaving (not making future purchases) based on their engagement data and loyalty points?

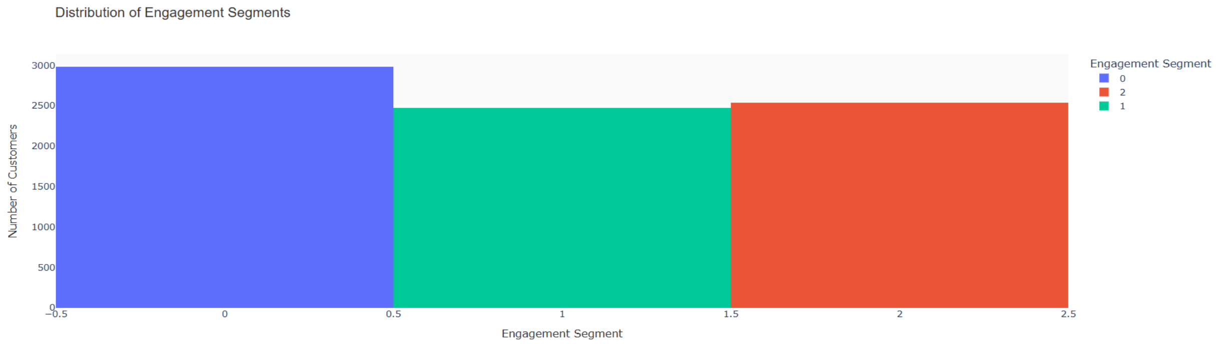
Answer: Number of at-risk customers: 4164 out of 8000 customers.



Problem2: How can we segment customers based on engagement behavior to develop targeted marketing strategies for each group?

Answer: Insights by Segments:

The engagement metrics across segments suggest tailored strategies to boost purchases. Segment 0 has high page views but low click-through and conversion rates, so enhancing content marketing and SEO could encourage more engagement. Segment 1 shows the highest click-through rate, indicating responsiveness to marketing efforts; targeted campaigns with strong Calls-to-Action (CTAs) may drive conversions here. Segment 2, with the highest conversion rate but lower page views, would benefit from personalized outreach and incentives to capitalize on purchasing readiness. These strategies can enhance segment-specific engagement and overall sales.



Problem 3: Conduct hypothesis testing to evaluate the effectiveness of Social Media and PPC (Pay-Per-Click) marketing campaigns. The goal is to determine if the conversion rates of both channels are equal, assessing whether they drive conversions equally effectively.

Answer:

Mean Conversion Rate (Social Media): 0.1066

Mean Conversion Rate (PPC): 0.1041

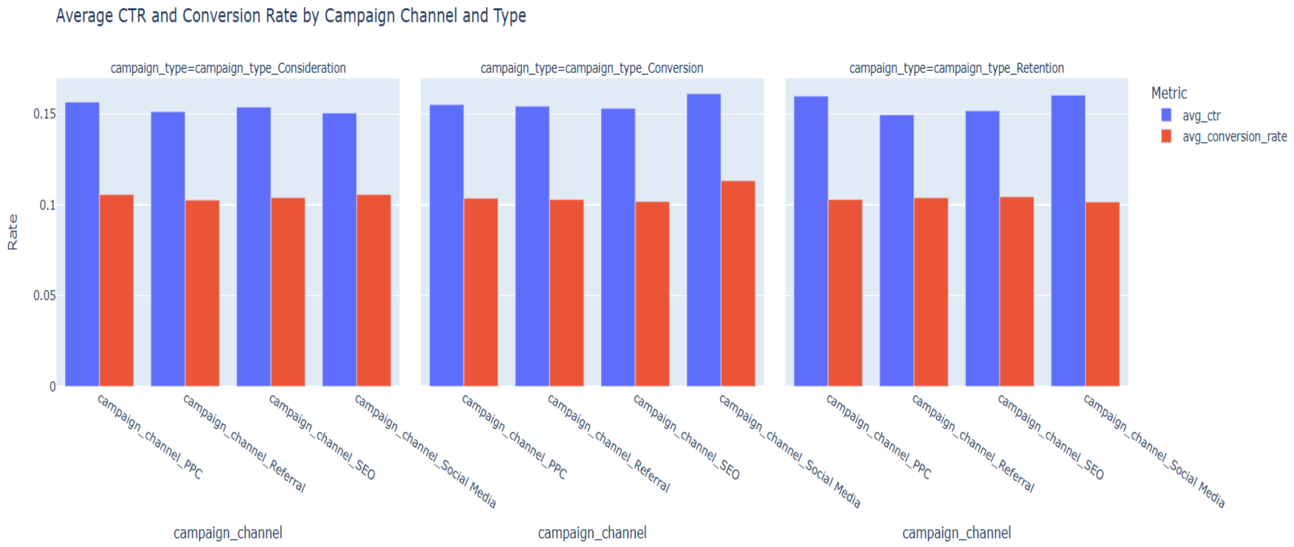
T-Statistic: 1.2402

P-Value: 0.2150

Fail to reject the null hypothesis: No significant difference in conversion rates between Social Media and PPC (Pay-Per-Click) campaigns.

Problem 4: How effective are different marketing channels and campaigns in driving customer conversion and engagement, as measured by CTR and conversion rate?

Answer: for conversion_through_rate and conversion_rate for both the best campaign_type is **Conversion** and campaign_channel is **Social Media** also can be seen at graph



Main Problem 5: How can we use classification models to predict customer conversion based on factors like demographics, campaign performance, and engagement metrics?

Answer: To solve this main problem firstly, we selected important features according to our targeted variable using models or methods: Random Forest Classifier and Gradient Boosting. After this we apply six algorithms and their Hyperparameter tuning using **Grid Search** method that are suitable for **Nonlinear data.**

Important Features for Predicting Conversion, Identified as Common across both methods mentioned above:

Feature Set (X):

- Total Page Views
- Time on Site
- Click-Through Rate
- Ad Spend
- Conversion Rate
- Loyalty Points
- Email Activity
- Income
- Previous Purchases
- Social Shares
- Age

Target Variable (y):

- conversion

Machine Learning Models used and their accuracies before Hyperparameter Tuning:

- **Decision Tree Accuracy:** 81.06%
- **Random Forest Performance Accuracy:** 89.31%
- **Gradient Boosting Performance Accuracy:** 89.94%

Models Oversampled for Class Imbalance:

To address poor performance on the minority class (class 0), oversampling was applied to models that showed significant imbalance. This technique aimed to improve recall and precision for class 0 by augmenting its sample size.

- **SVM (RBF Kernel) Accuracy**: 54.69%
- **K-Nearest Neighbors Accuracy**: 62.88%

This Deep learning model had also issue of poor performance with minority class (class 0):

- **Neural Network (MLP) Accuracy:** 17.81%

Results of models after Hyperparameter Tuning using (Grid Search Method):

Model	Before Tuning Accuracy	After Tuning Accuracy	Difference	Improvement	Best Parameters
Gradient Boosting	**89.94%**	**90.38%**	**+0.44%**	**Best Model**	**{'learning_rate': 0.1, 'max_depth': 3, 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 100, 'subsample': 0.8}**
Random Forest	**89.31%**	**89.19%**	**-0.12%**	**Second Best Model**	**{'bootstrap': False, 'max_depth': None, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 100}**
Decision Tree	**81.06%**	**88.50%**	**+7.44%**	**Moderate Improvement**	**{'max_depth': 3, 'max_features': None, 'min_samples_leaf': 1, 'min_samples_split': 2}**

Neural Network (MLP)	**17.81%**	**87.88%**	**+70.06%**	**Significant Improvement**	**{'activation': 'relu', 'alpha': 0.0001, 'hidden_layer_sizes': (50,), 'learning_rate': 'constant', 'solver': 'sgd'}**
SVM (RBF Kernel)	**54.69%**	**87.88%**	**+33.19%**	**Good Improvement**	**{'C': 0.01, 'gamma': 'scale', 'kernel': 'rbf'}**
K-Nearest Neighbors	**62.88%**	**87.75%**	**+24.87%**	**Moderate Improvement**	**{'metric': 'euclidean', 'n_neighbors': 11, 'weights': 'uniform'}**

Report Summary:

This project analyzes customer engagement and loyalty data to predict conversion likelihood, segment customers for targeted marketing, and assess marketing channel effectiveness. Through classification models and hypothesis testing, insights were derived on customer risk, channel performance, and conversion drivers.

Future work:

Future work for this project includes developing a Streamlit-based web app for real-time prediction of conversion likelihood, where users can input customer demographic and engagement data, and the app will display prediction probabilities and customer segmentation using the best classification model.

Another area for improvement is applying regression models, such as Linear Regression and Gradient Boosting Regressor, to predict the conversion rate with greater precision. Performance evaluation using metrics like RMSE and MAE will refine these models further.

Lastly, deep learning models could be explored for complex data patterns. Neural network architectures in TensorFlow or PyTorch, with hyperparameter tuning and optimizers like Adam, could enhance prediction accuracy compared to traditional machine learning methods.

Acknowledgement:

We sincerely appreciate Miss **Reema** for her invaluable support throughout this project. Her guidance, insightful notes, and resourceful links greatly contributed to the successful completion of this work. This project would not have been possible without her dedicated assistance 🙏 and the collaborative efforts of our team🤝.

A sincere thank you to **reviewers, Gexton** and **PITP** management for their ongoing support and encouragement. We are also grateful to the **Kaggle community** for providing access to high-quality, accurate data, which played a crucial role in our analysis.