Predictive Marketing Analytics for Better Lead Conversion in EdTech

A Mid-Term report for the BDM Capstone Project

Submitted by

Name: AKHILESHWAR PANDEY

Roll number: 21F3002866

IITM Online BS Degree Program,

Indian Institute of Technology, Madras,

Chennai Tamil Nadu, India, 600036

Contents

Executive Summary

Blade Learners is an EdTech company that provides study materials, test series, and answer-writing assistance for humanities students preparing for board exams and undergraduate entrance tests. The major challenge the company faces is a low lead conversion rate of about 7 percent. which means that despite reaching out to a large number of potential students, very few actually enroll in the courses.

The key issues identified in the sales process are:

- Lack of prioritization in calling leads, leading to inefficient resource use.
- No structured way to predict which leads are most likely to convert.
- A high drop-off rate in follow-up calls, reducing engagement.

To tackle these problems, a data-driven approach was taken:

- Understanding the sales process and defining objectives.
- Collecting and analysing past lead data to identify trends.
- Using Excel for preliminary data processing and then shifting to Python for deeper analysis.
- Visualizing key insights to understand conversion patterns.
- Building a predictive model to score leads based on their likelihood of conversion.

A detailed breakdown of the dataset (stored in Excel format) is provided under the Metadata section. The data includes lead details, course preferences, call status, and conversion outcomes.

Further Descriptive statistics is obtained to identify key patterns in data and get some firsthand insights.

Visual representations were created to better understand trends in lead conversion and engagement patterns.

With this foundation, the next steps involve using data analysis techniques to refine sales strategies and develop a predictive model to support decision-making

Proof of Originality of the Data

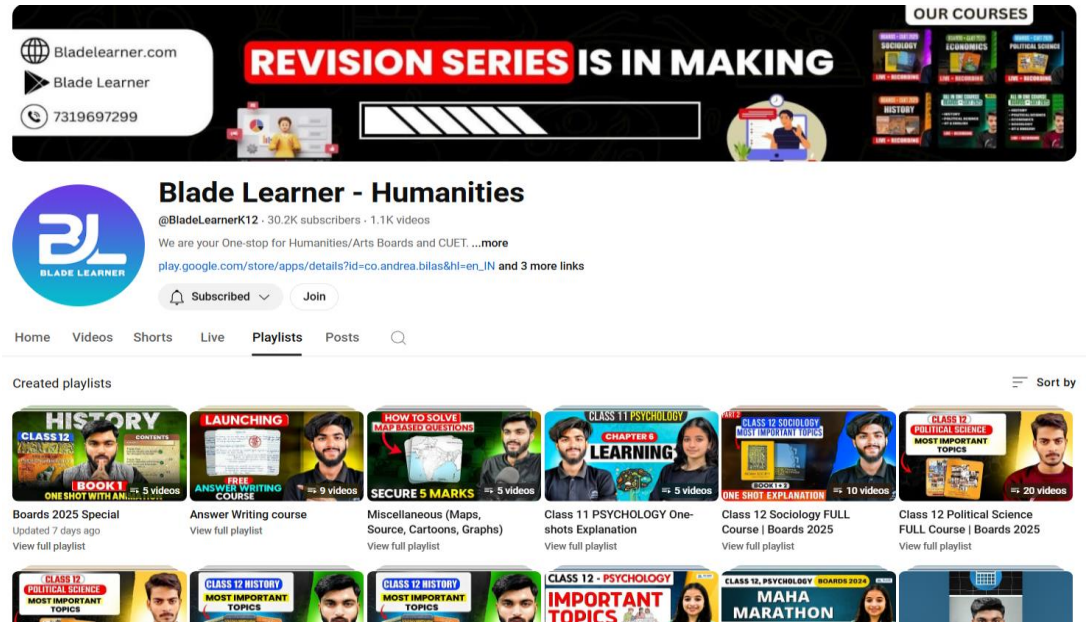
Business Name: Blade learner Pvt. Ltd.

Address: Noida, Uttar Pradesh (Delhi NCR)

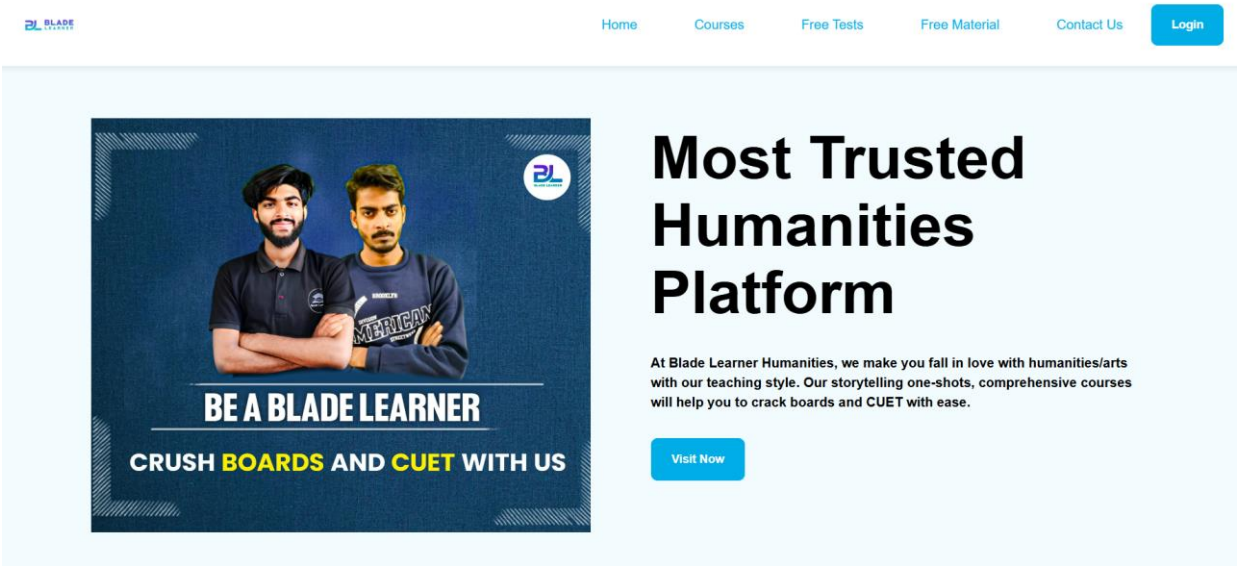
Founders: Sudhanshu Kumar, Ayush Raj, Faran Alam

CIN No: U80903BR2022PTC059116

Video of Interaction with Sales Head (Blade Learners) | Field Notes



Blade Learner, You Tube Page



Blade Learner Web Application

Metadata

BDM Project Data: Data

Data Format: Excel/Sheets (XLSX)

Range: October 1,2024, to January 5, 2025

Column Name	Data Type	Description	Unique Values
ID	Integer	Unique ID for each row	Continuous Numeric Value
Lead Type	String	Category of Lead Based on its Source	NORMAL, App Query, WARM, HOT, WhatsApp, App Download, YouTube, Website Query, Instagram
Clicked Course	String	Course Clicked on app/web or showed interest	Alpha All in One, Others, Unknown, Alpha, Eco Ninja, His Ninja, Alpha OG, Socio Alpha, His Alpha, Pol Ninja, Ninja, Pol Alpha, Eng + GT Alpha
Date	Datetime	Date of registering Lead	Datetime Values
1st Calling Date	Datetime	Date of first call	Datetime Values
Diff1	Integer	Days between first call and lead registered	Continuous Numeric Value
Class	String	Class/Standard of student	11, 12, Unknown
Subjects	String	Subject(s) student interested.	'Unknown', 'His, Pol, Geo, Eng', 'Others', 'His, Pol, Socio, Eng', 'His, Pol, Eco, Eng', 'His,Pol,Geo,Eng', 'His,Pol,Eco,Eng', 'His, Pol, Geo', 'His, Pol, Eco'
1st Call/ 2nd Call/ 3rd Call	String	Status of First, Second or Third Call.	Unknown, TD, NR, CB, CC
Probability1/ Probability2/ Probanility3	String	Degree of interest of student as assessed on first, second or third call.	Unknown, Hurray Won! High Chances (>75%), Not Interested, Chances (40-75%), Less Chance (<40%), Connected(in Loop), Not Interested in Talk
Diff2	Integer	Days between first and second call.	Continuous Numeric Value
3rd Calling Date	Datetime	Date of Third Call	Datetime Values
Diff3	Integer	Days between second and third call.	Continuous Numeric Value
Parents	Integer	Whether parents' intervention was sought by student.	0,1
Origin	String	Whether Lead is organic or inorganic	Organic, Inorganic
Target	String	Final Outcome: whether lead was won or lost.	Won, Lost

Descriptive Statistics

Following Statistics is about the count of each unique Values in Categorical Columns and relevant Numerical Columns. Ratio is count divided by total number of rows.

Feature: 1st Call

1st Call	Count	Ratio
Unknown	260	0.472727
TD	135	0.245455
NR	100	0.181818
CB	21	0.038182
CC	19	0.034545

Feature: Clicked Course

Clicked Course	Count	Ratio
Alpha OG	116	0.210909
Others	90	0.163636
Ninja	53	0.096364
Alpha	49	0.089091
His Ninja	45	0.081818
Alpha All in One	37	0.067273
His Alpha	36	0.065455
Eng + GT Alpha	28	0.050909
Socio Alpha	27	0.049091
Unknown	23	0.041818
Pol Ninja	23	0.041818
Eco Ninja	12	0.021818
Pol Alpha	11	0.020000

Feature: Lead Type

Lead Type	Count	Ratio
NORMAL	420	0.763636
WARM	54	0.098182
App Query	35	0.063636
HOT	12	0.021818
WhatsApp	9	0.016364
YouTube	8	0.014545
App Download	5	0.009091
Instagram	3	0.005455
Already Enrolled	2	0.003636
Website Query	2	0.003636

Feature: Origin

Origin	Count	Ratio
Inorganic	532	0.967273
Organic	18	0.032727

Feature: Probability1

Probability1	Count	Ratio
Unknown	405	0.736364
High Chances(>75%)	33	0.060000
Not Interested	33	0.060000
Chances(40-75%)	26	0.047273
Hurray Won!	17	0.030909
Connected(inLoop)	14	0.025455
Less Chance(<40%)	12	0.021818
Not Interested in Talk	10	0.018182

Feature: 2nd Call

2nd Call	Count	Ratio
U	482	0.876364
NR	29	0.052727
TD	25	0.045455
CB	5	0.009091
CC	5	0.009091

Feature: 3rd Call

3rd Call	Count	Ratio
U	539	0.980000
TD	5	0.009091
NR	4	0.007273
CB	2	0.003636

Feature: Target

Target	Count	Ratio
Lost	510	0.927273
Won	40	0.072727

```
Feature: Probability2
                       Count     Ratio
Probability2
Not Interested            10  0.018182
High Chances(>75%)         7  0.012727
Chances(40-75%)            5  0.009091
Hurray Won!                3  0.005455
Not Interested in Talk     3  0.005455
Less Chance(<40%)          2  0.003636
Connected(inLoop)          1  0.001818
```

Metric	Diff1	Diff2	Diff3
Valid Entries	297.0	69.0	11.0
Mean Days	1.0	5.0	7.1
Std Deviation	2.3	5.2	5.4
Minimum	0.0	0.0	2.0
25th Percentile	0.0	1.0	2.0
Median	0.0	3.0	7.0
75th Percentile	1.0	6.0	10.5
Maximum	25.0	18.0	16.0

```
Feature: Subjects
                     Count     Ratio
Subjects
Unknown                416  0.756364
Others                  59  0.107273
His, Pol, Geo, Eng      27  0.049091
His, Pol, Eco, Eng      24  0.043636
His, Pol, Socio, Eng     8  0.014545
His, Pol, Geo            6  0.010909
His, Pol, Eco            4  0.007273
His,Pol,Geo,Eng          3  0.005455
His,Pol,Eco,Eng          3  0.005455
Feature: Probability 3
                       Count     Ratio
Probability 3
Not Interested             2  0.003636
Hurray Won!                1  0.001818
High Chances(>75%)         1  0.001818
Purchased another Batch    1  0.001818
Chances(40-75%)            1  0.001818
```

- 25% of leads receive first call same-day (Diff1 25th percentile = 0)
- Median response time doubles with each follow-up (1 → 2 → 4 days)
- Third-call intervals show highest variability (Std Dev 4.3 days)
- Extreme cases show 2-week delays in follow-ups (Max 14-17 days)

- The majority of leads (76.4%) are classified as "NORMAL," followed by "WARM" (9.8%).
- Organic sources such as WhatsApp, YouTube, Instagram, and Website Queries contribute minimally to the overall lead pool.
- "Alpha OG" is the most clicked course (21.1%), while "Others" collectively account for 16.4%.
- 12th-grade students form the largest identifiable segment (24.5%).
- The most common subject combinations include "His, Pol, Geo, Eng" (4.9%) and "His, Pol, Eco, Eng" (4.4%).

- Out all 530 leads registered calls were made to only 283 students in first round, of these only 66 were called for second time and further only 35 were called in third round.

- Of all the 283 students called in first round about 52% student Talked.
- Of all the 66 students called in second round about 38% student Talked.
- Of all the 35 students called in third round about 7% student Talked.
- 47.3% of leads remain "Unknown" in the first call status, suggesting follow-up inefficiencies.
- Inorganic leads dominate the dataset (96.7%) and organic leads constitute only 3.3%.
- Of all 550 leads registered only 40 could be converted that gives 7 percent conversion rate.

Detailed Explanation of Analysis process

The organization struggles to identify high-potential leads, resulting in an unstructured tele-calling approach and low conversion rates. It also aims to evaluate follow-up timelines and frequencies to improve engagement.

Objective

Main objective of this work is to analyse lead data, derive meaningful insights, and assess the sales funnel to identify inefficiencies in the engagement strategy. By exploring conversion patterns and follow-up interactions, the objective is to establish a data-driven approach to optimize tele-calling efforts. Based on these insights, a predictive classification model will be built to prioritize high-potential leads, enabling better resource allocation and improved conversion rates.

Approach adopted to move ahead in this project was firstly getting intuition about dataset using MS excel do some preliminary processing there and then move to python for further preprocessing of Data and Exploratory data Analysis. After that some feature engineering would be done to prepare data for building classification ML model to predict the ‘Target’ column.

Preliminary Preprocessing (MS Excel)

After getting the data sheet of leads, preliminary examination and processing of data was performed on MS Excel as it provides a quick, intuitive, and efficient way to inspect, clean, and manipulate raw data before further analysis using Python.

- In excel filter function inside Data menu was used extensively to filter and group the data which proved very useful.
- Originally there were 570 rows among which 550 were retained as rest of them have a lot of discrepancies.
- Blank columns were filled with appropriate keywords such as Unknown etc.
- New columns Diff1, Diff2, Diff3 were created to calculate the number of days between successive calls.
- Further Remarks column which contained subjective remarks which can't be much helpful in any type of analysis was discarded after seeking information whether students sought their parents advise and new column 'Parents' being created.

Preprocessing and Exploratory Data Analysis (Python)

After these preprocessing steps, data was imported to Google Colab for further preprocessing and exploration, as Python provides powerful libraries such as Pandas, Matplotlib, Seaborn, Scikit-Learn etc for efficient data handling, transformation, and analysis. Google Colab, specifically, was chosen due to its cloud-based environment, which eliminates the need for local computational resources.

- Data was stored as Pandas dataframe df, after that basic information like about data was sought using head (gives first five rows), info (basic info of data), describe (descriptive statistics of numerical columns), isnull (number of null values) etc functions of pandas library.
- Number of unique values in each categorical column was found using unique and nunique function. Columns 'Clicked Course' and 'Subjects' had 65 and 60 unique values respectively. In these columns unique values having less frequent occurrences were grouped together as new unique value 'Others' in order to avoid the clutter.
- Descriptive Statistics presented in previous section that is frequency of each unique values in categorical columns and their percentage with respect to total number of observations was calculated in order to get the overview of different values in columns.

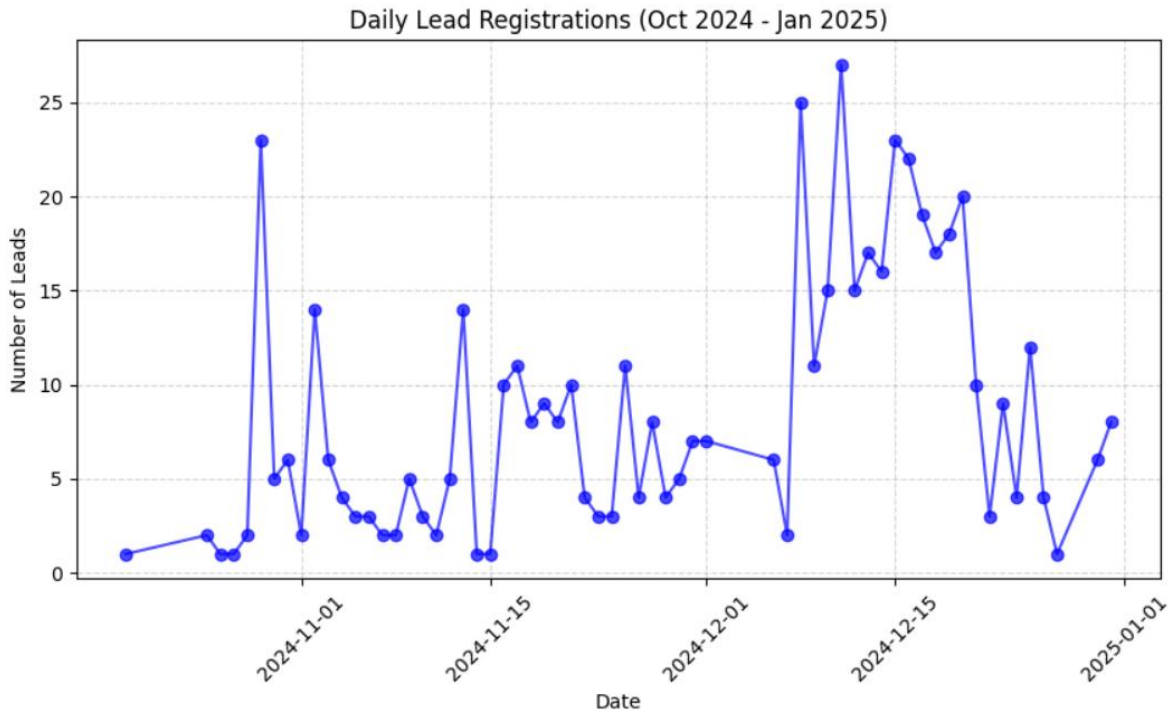
- Python Matplotlib and seaborn libraries were used to make different graphs and charts to get pictorial insights of the data.
- Conversion rate with respect to Lead Type and Clicked Course was plotted to get the insight about how these variables affect the conversion.

ML Classification Model Building

- After this Some feature Engineering like encoding of categorical features, generating polynomial features etc, would be done on data to make it suited for building ML classification model.
- Data would be splitted in 80:20 ratio for training and validation.
- Then a baseline model, most probably which is a logistic regression model would be built and it would be fine-tuned using hyperparameter tuning libraries. For these purposes libraries like scikit-learn, train-test-split etc would be used.
- Using the test data model is evaluated using metrices like Precision, Recall, F1-score.

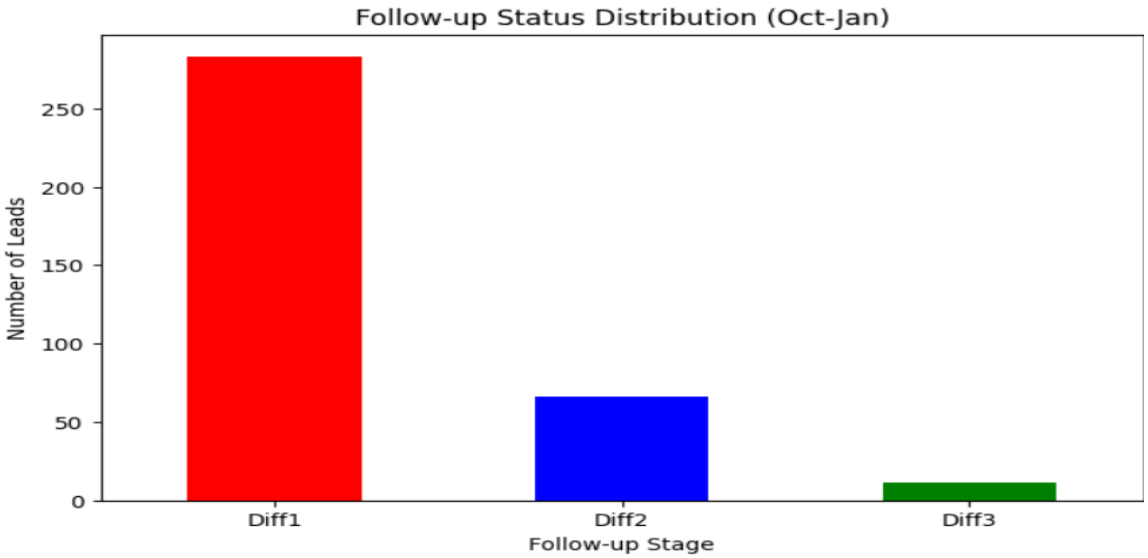
Results and Findings

Some of the results and findings are listed below:



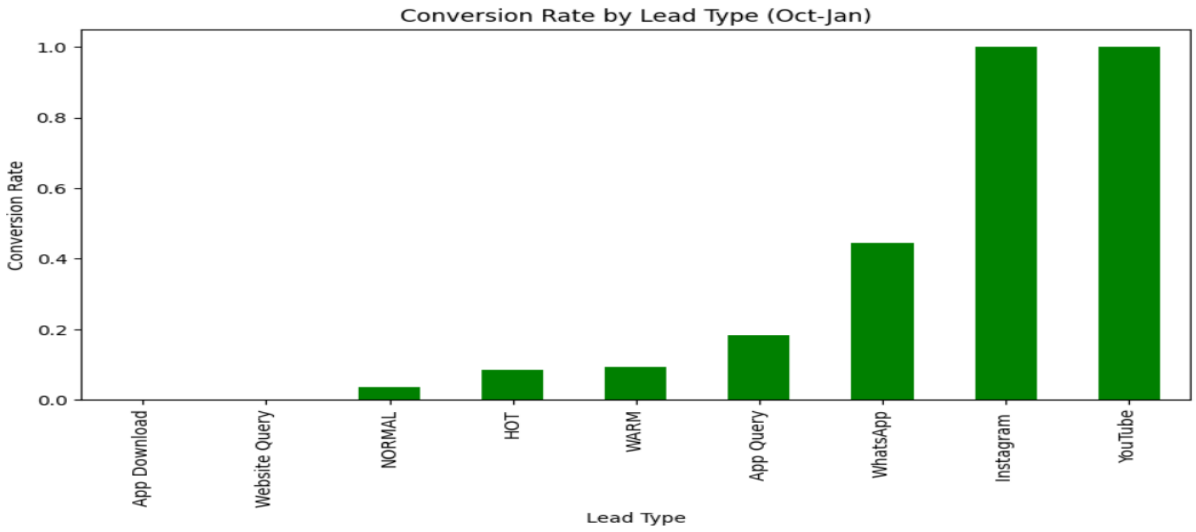
Graph 1: Daily Lead Registration

- The number of daily lead registrations varies significantly throughout the period, with noticeable peaks and troughs, but there is a clear surge in daily lead registrations during December 2024, specifically in mid-December.



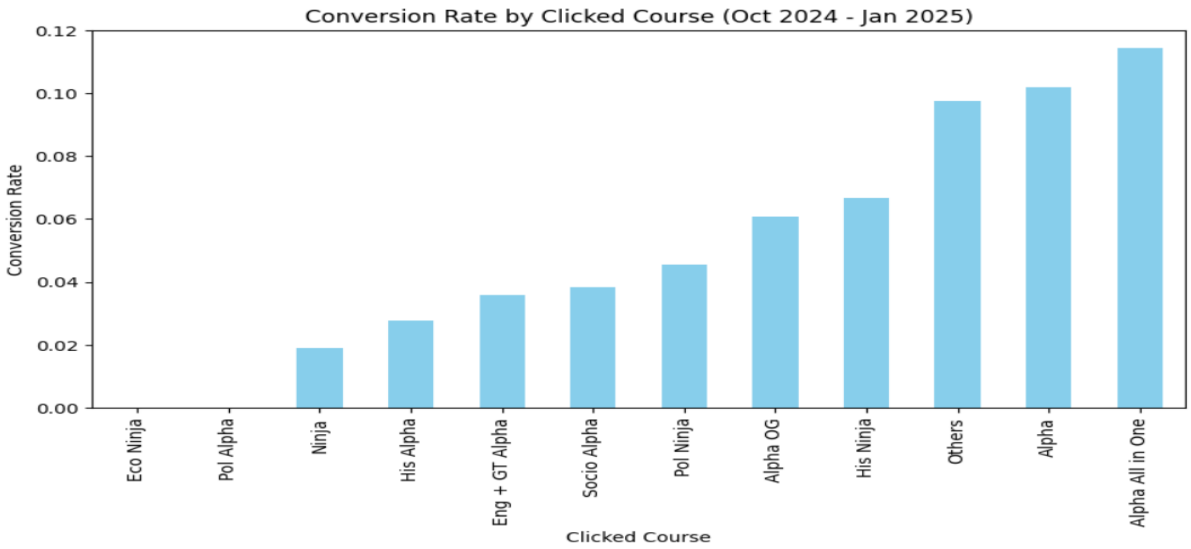
Graph2: Follow-up Distribution

- Above is distribution of follow-up which shows that with each subsequent level number of calls decreases by more than half.



Graph 3: Conversion Rate by Lead Type

- Although leads from WhatsApp, Instagram and YouTube are less in volume but their conversion rate is much higher as compared to other leads. Also, 'NORMAL' which happens to be the most frequent lead type is having very less conversion rate.



Graph 4: conversion Rate by Clicked course

- The graph above shows that 'Alpha All in One' has highest Conversion rate. We know that Alpha OG has highest frequency but its conversion rate is lower than three courses.