Lead Scoring Case study

- By Anil Kumar M, Tuyet tu, Anil Goud Gunda

Business Objective & Problem Statement

Business Objective & Problem Statement

Business Context: X Education is an online education provider that attracts leads through various marketing channels.

Problem Statement: The current lead conversion rate is low (30% on average).

Objective: Develop a logistic regression model that assigns a lead score (0–100) to each prospect, helping identify “hot leads” and aiming to boost the conversion rate toward 80%.

Exploratory Data Analysis (EDA)

Data Overview

Dataset Details: 9,240 observations and 37 variables including customer demographics, engagement metrics (e.g., Total Visits, Time Spent on Website), lead quality, and more.

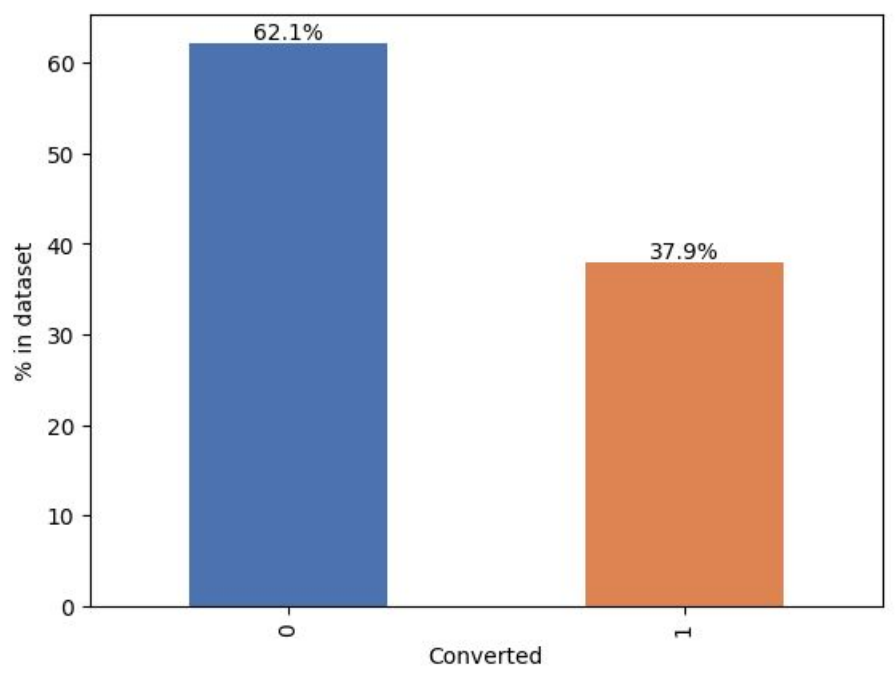
Key Variables: Prospect ID, Lead Origin, Lead Source, Converted, TotalVisits, Total Time Spent on Website, Page Views Per Visit, Lead Quality, Tags, etc.

Data Challenges: Presence of missing values and scattered categories (e.g., in Lead Quality and Tags).

Data Imbalance

37.9% of the 'Converted' data is 1 ie. 37.9% of the leads are converted. Generally, this is a moderate imbalance dataset.

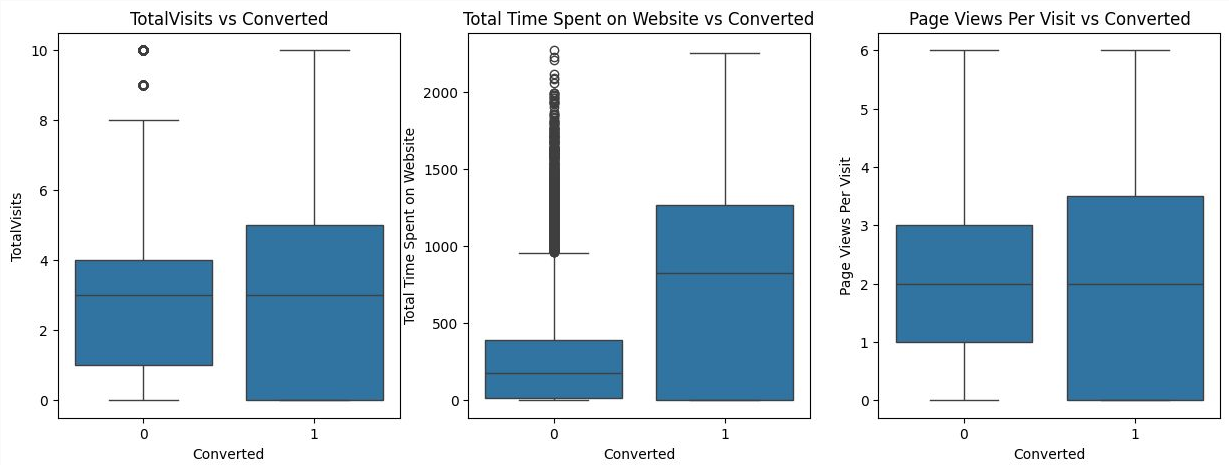
Logistic Regression can handle this level of imbalance



Numeric variable vs Target variable

Observations:

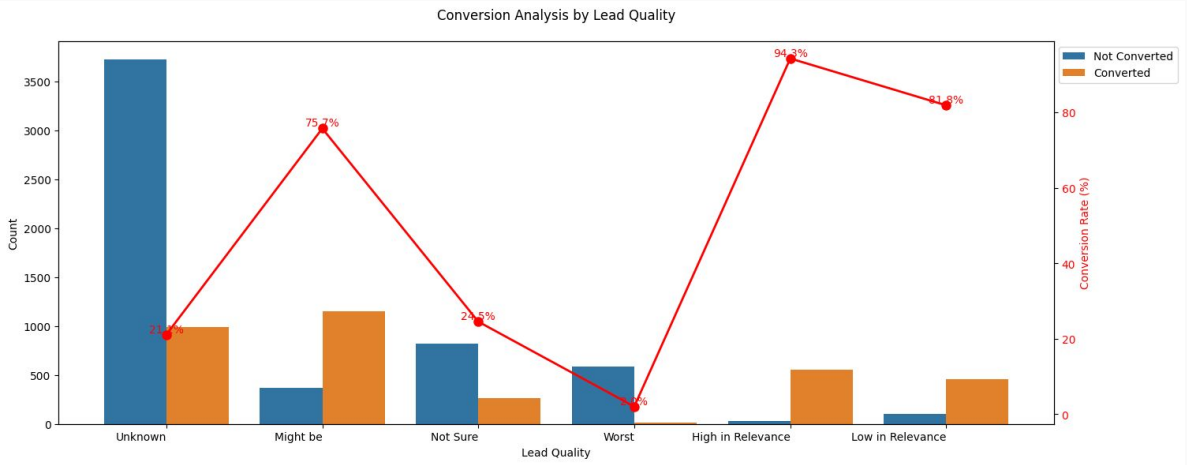
'TotalVisits' has same median values for both groups. However, converted users are more spreading.
Converted leads spent significantly higher time on site with a higher median than non-converted group. We also see a larger spreading and fewer outliers in converted group, meaning a certain range of time on site is optimal for converting with certain variance of behavior in this group
'Page Views Per Visit' also has same median values for both groups.
=> We can see that time spent appearing to be a good predictors of conversion



Conversion Analysis by Lead Quality

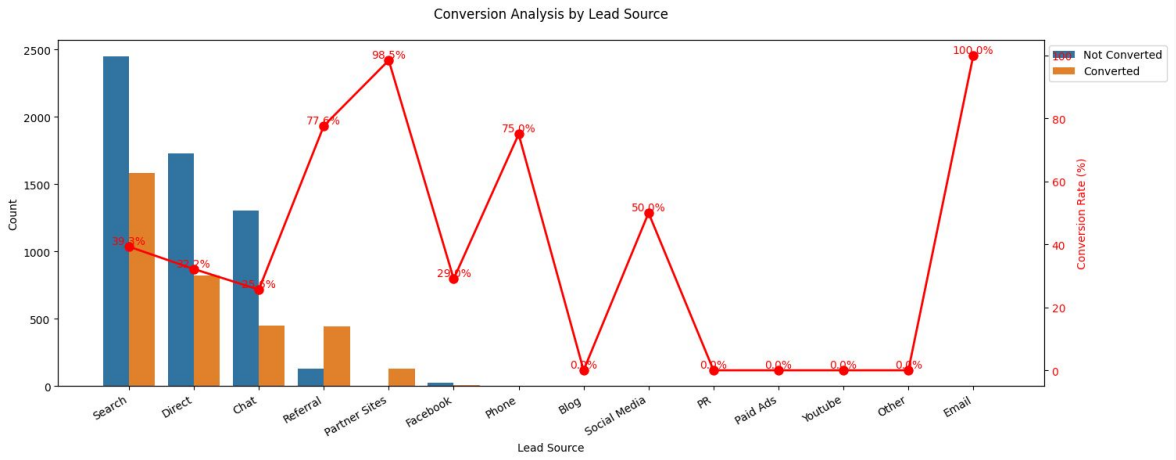
"High in Relevance" and "Low in Relevance" leads have the highest conversion rate (~94%), however, they have relatively low volume.

"Unknown" has a high volume, but low conversion rate, suggesting that we should focus on this group for lead targeting
"Might be" show a strong performance and good balance of volume and conversion rate



Conversion Analysis by Lead Source

High volume sources: search, direct, and chat has most traffic, yet low conversion rate
Referral marketing from parter sites, referral programs shows a strong conversion rate because they are more trusted sources
Digital marketing channels, except Social Media and Facebook, have a very poor result

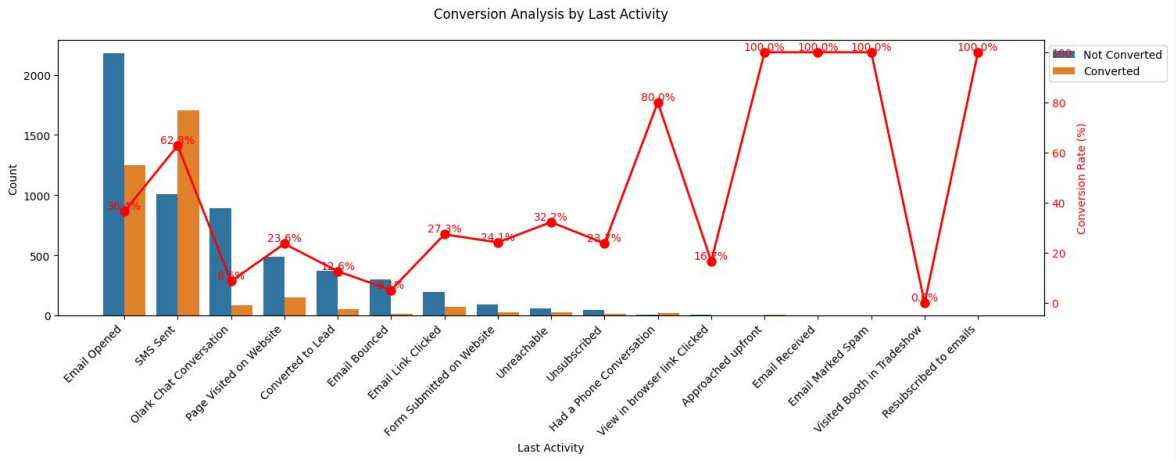


Conversion Analysis by Last Activity

Highest number of lead are generated where the last activity is 'Email Opened', 'SMS Sent' and 'Chat Conversation', out of which, SMS is better conversion rate.

Others have too low volume to evaluate the conversion rate properly

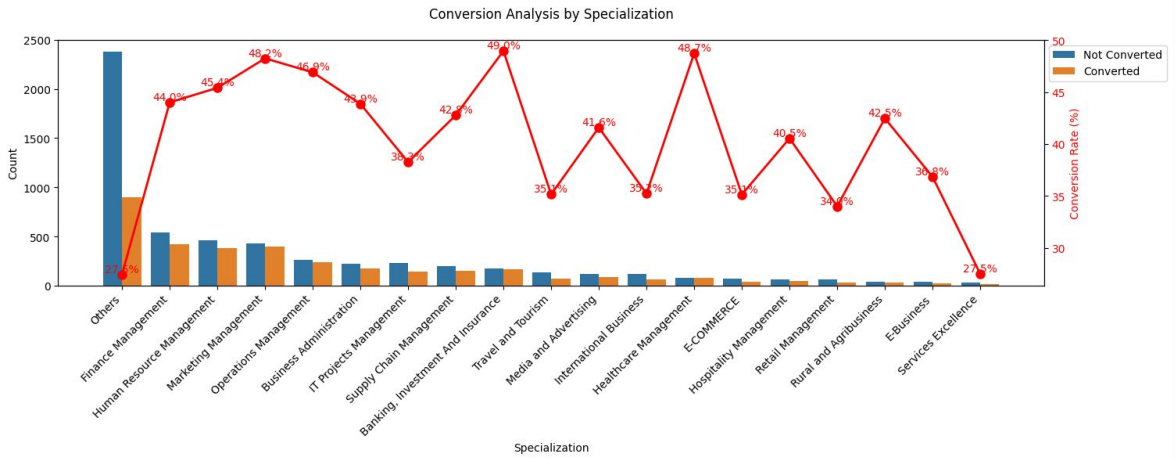
"Page Visited on website" is a good indicator for early stage interaction with 23% converted



Conversion Analysis by Specialization

"Others" has highest volume but lowest conversion rate while most specialised fields have relatively low volumes, but higher conversion rates

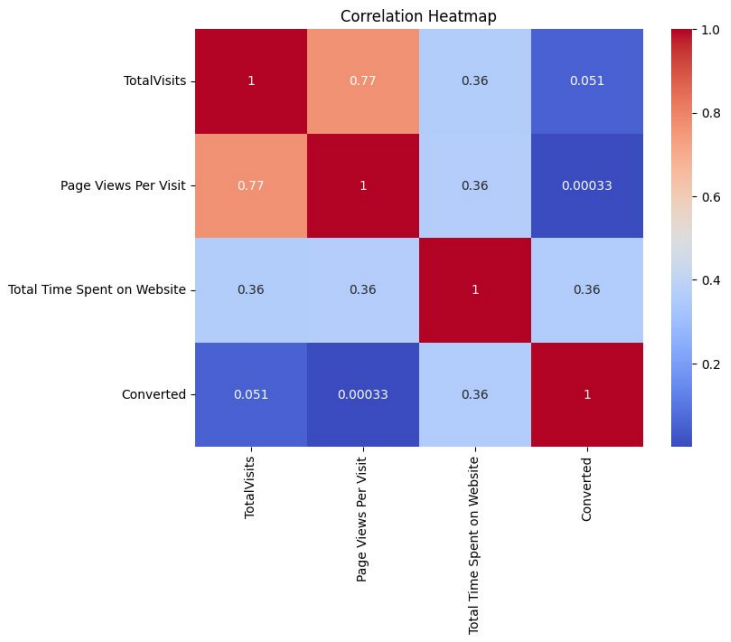
Suggesting that if we can develope strategies to categorize these leads into specific specializations, helping us to improve overall conversion by better classification



Correlation Heatmap (Numeric Features)

Strong correlation between Total Visits and Page Views per visit, suggesting that people who visits more frequently also tend to view more pages per visit

Again, time on site has a moderate correlation with "Converted", a good predictive variable for conversion



Data Preprocessing

Data Understanding & Cleaning

1. Inspected data types, dimensions, and explored initial dataframe (using df.head(), df.info(), df.describe()).
2. Columns with >20% missing values (Lead Quality, Tags, etc.) were imputed or strategically handled (e.g., 'Lead Quality' imputed as 'Unknown', 'Tags' grouped and imputed).
3. Numerical columns with <20% missing values imputed with median.
4. Categorical columns with <20% missing values imputed with 'Unknown'.

Feature Engineering & Transformation

1. Categorical feature grouping (e.g., 'Tags', 'Occupation', 'Asymmetrique Activity Index') to simplify categories and improve interpretability.
2. Dropped less informative columns (e.g., 'Country', 'How did you hear about X Education', 'What matters most to you in choosing a course', 'Lead Profile').
3. Converted binary categorical features ('Yes/No') to numerical (1/0).
4. Encoded remaining categorical features using Label Encoding.

Data Scaling & Splitting

For numerical features, specifically 'TotalVisits', 'Page Views Per Visit', and 'Total Time Spent on Website', were scaled using MinMaxScaler. to transforms each feature to a range (between zero and one), mitigating the influence of features with inherently larger scales and ensuring equitable contribution to the model.

Then we partitioned the data into training (80%) and testing (20%) subsets. To model training and test the final model performance using test data set

Model Building & Feature Selection

Feature Selection using

Recursive Feature Elimination (RFE)

We followed Recursive Feature Elimination (RFE) to systematically identifies and ranks features selecting the top 15 features through RFE,

The key features identified through this process notably included elements such as Lead Origin, Lead Quality, Tags, and Total Time Spent on Website, representing critical factors in lead conversion prediction.

Iterative Model Building, RFE & VIF Analysis

Model Building Process:

We Followed an iterative approach to refine our Logistic Regression model, starting with features selected through Recursive Feature Elimination (RFE). To address multicollinearity and improve model stability, we utilized Variance Inflation Factor (VIF) analysis, systematically removing features exhibiting high collinearity, such as 'What is your current occupation', 'Do Not Call', and others.

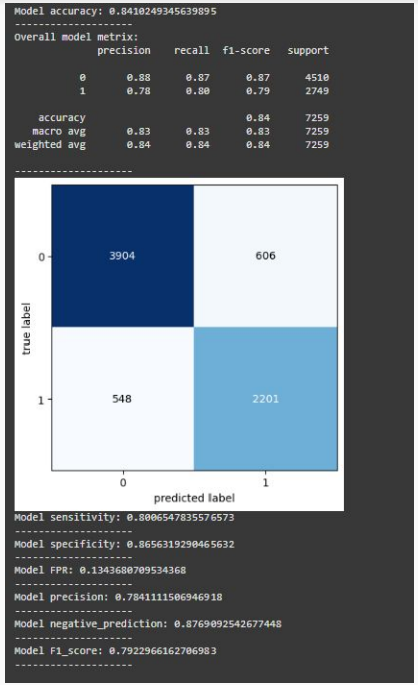
With this process we identified key features including Lead Origin, Lead Source, Do Not Email, Total Time Spent on Website, Page Views Per Visit, Tags, Lead Quality, and Last Notable Activity, among others for a better accurate model

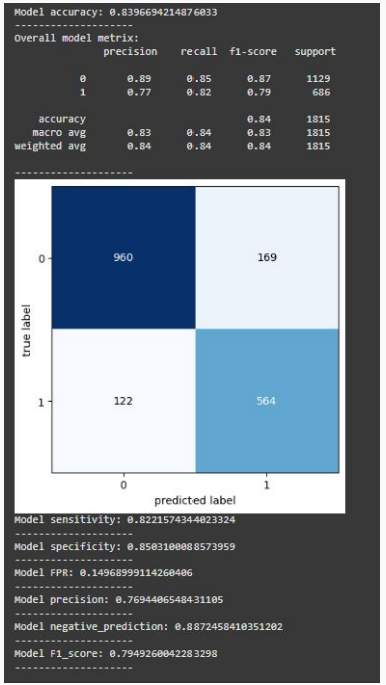
Model Evaluation & Performance

Model Evaluation Metrics

Model shows outstanding stability with nearly identical performance between training (84.10%) showing better performance on Non-converted leads (Class 0: f1-score 87% compared to 79% of Class 1)) and strong ability to avoid false positives.

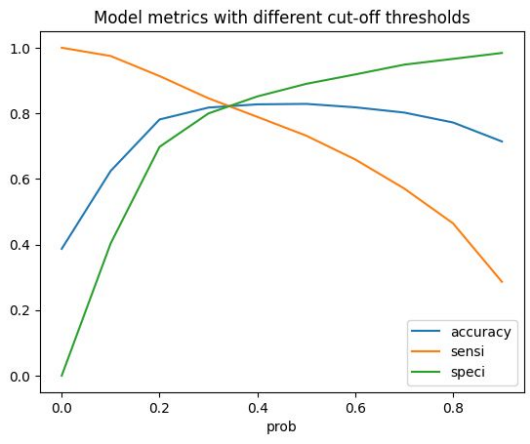
and The model performs consistently on the test set with the accuracy of 83.9%

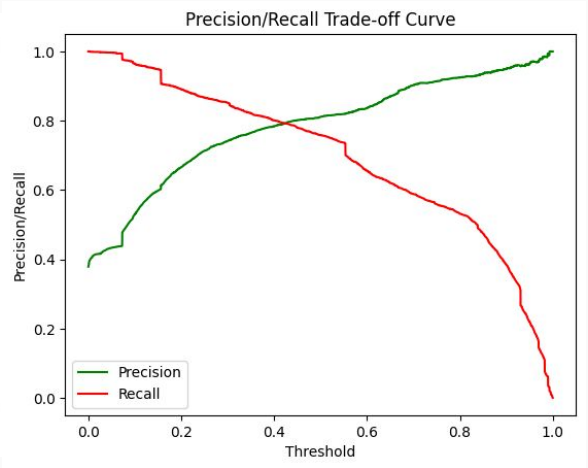




Optimal Cut-off & Confusion Matrix

At a 0.4 cut-off, the model achieves a good balance between correctly identifying potential leads and minimizing false positives.



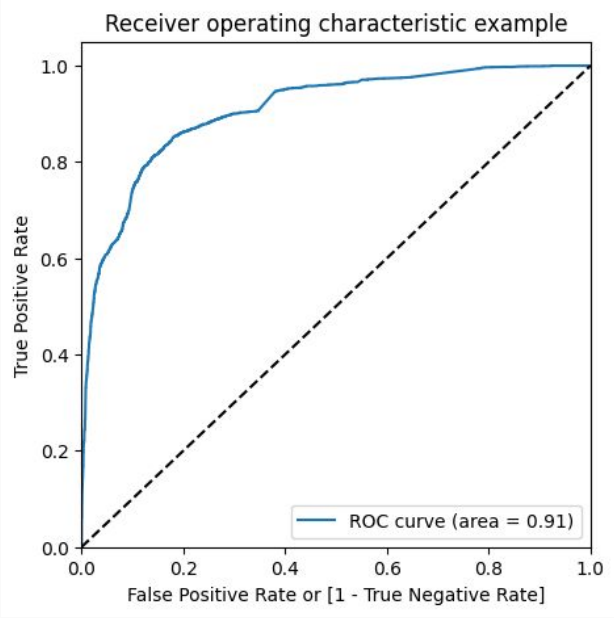


ROC Curve

Model Performance:

AUC = 0.91 indicates excellent discriminative ability of the model

Curve far above diagonal = model significantly outperforms random classification



Top Features

Top Features

Key Positive Predictors (increasing conversion probability): Lead Source_Partner Sites (3.86): Highest positive coefficient, as we can see from the above analysis, partnership is a reliable channel that gives more trust and credibility to users to enrol in the courses

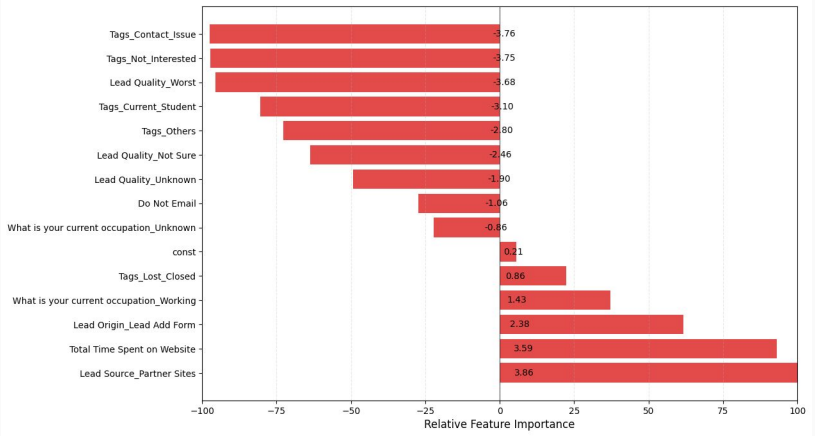
Total Time Spent on Website (3.59): Strong positive impact. The model fits with our EDA insights which suggest a strong correlation between the conversion rate average time the users spend on the website

Lead Origin_Lead Add Form (2.38): Strong positive influence, clearly, filling the form is a strong indicator of the users' interest in the courses

Key Negative Predictors (decreasing conversion probability): Tags_Contact_Issue (-3.76): Strong negative effect. Not being able to contact the leads, is a strong indicator that they don't want to be further communicated about the courses

Tags_Not_Interested (-3.75): Strong negative influence. As tagged by employees, users strongly expressed the disinterest in the courses

Lead Quality_Worst (-3.68): Strong negative impact. This lead quality criteria shows relevancy to the conversion quality



Recommendations

Recommendations

Boost Website Engagement: Prioritize website enhancements that encourage longer visits and deeper content exploration to capture higher intent leads.

Refine Lead Quality Scoring: Standardize and improve the lead quality assessment process to accurately identify and prioritize high-potential leads for focused outreach.

Optimize Key Lead Channels: Concentrate marketing efforts and resources on Google and Direct Traffic channels to maximize acquisition from these high-yield sources.

Personalize Lead Communication: Tailor communication strategies and content based on observed lead behavior and specialization interests to enhance relevance and drive engagement.