Lead Scoring Case Study

Dheeraj T, Chaithanya Modupalli, Kalai Arun

Goals of the Case Study

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Business Understanding

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals.

Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Problem Statement

The lead conversion rate at X Education is quite low.

For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To improve this process, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.

By successfully pinpointing these leads, the conversion rate is expected to increase as the sales team will concentrate their efforts on engaging with potential leads rather than reaching out to everyone.

Dataset

- The dataset provided is ‘Leads.csv’ which contains all the information about the leads.
- Each lead is provided a unique ‘Prospect ID’ and ‘Lead Number’.
- There are a total of 37 columns which provides data on Lead Origin/Source, notification preferences, Website statistics of the lead, Activities of the lead and certain indexes.
- It has a total of 9240 records.
- The target variable is a column named ‘Converted’ which has values ‘1’ and ‘0’ which represents ‘True’ and ‘False’ respectively.

Data Cleaning

Columns with more than 40% null values

- Lead Quality – 51.59% null values
- Asymmetrique Activity Index – 45.65% null values
- Asymmetrique Profile Index – 45.65% null values
- Asymmetrique Activity Score – 45.65% null values
- Asymmetrique Profile Score – 45.65% null values

Lead Quality column null values are replaced with ‘Unknown’ as it seems to be an important column.

The remaining three columns are dropped

Data Cleaning

Dropped the below columns

- Tags - 36.29% null values
- Country - 26.63% null values
- Lead Profile - 29.32% null values
- What is your current occupation - 29.11% null values
- How did you hear about X Education - 23.89% null values
- Specialization - 15.56% null values
- City - 15.37% null values

Replaced the below columns with mode/median

- What matters most to you in choosing a course - 29.32% null values - Replaced with mode value
- Page Views Per Visit - 1.48% null values - Replaced with median value
- TotalVisits - 1.48% null values - Replaced with median value
- Lead Source - 29.32% null values - Replaced with mode value

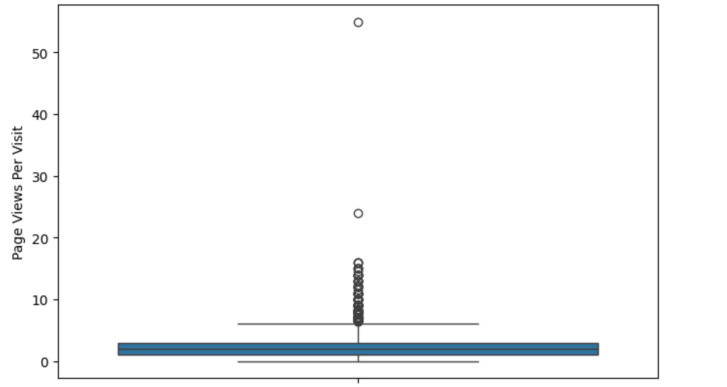
Removed the rows with null values

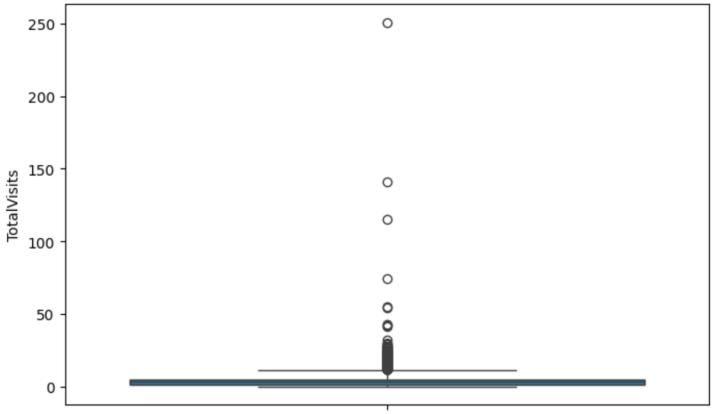
- Last Activity - 1.11% null values

Data Cleaning

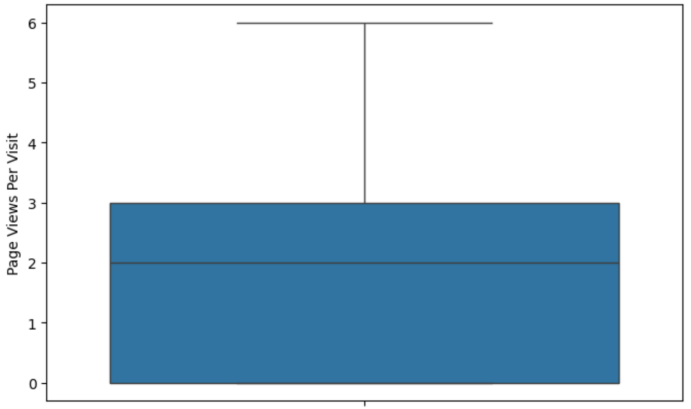
Removed Outliers of Page Views Per Visit and TotalVisits

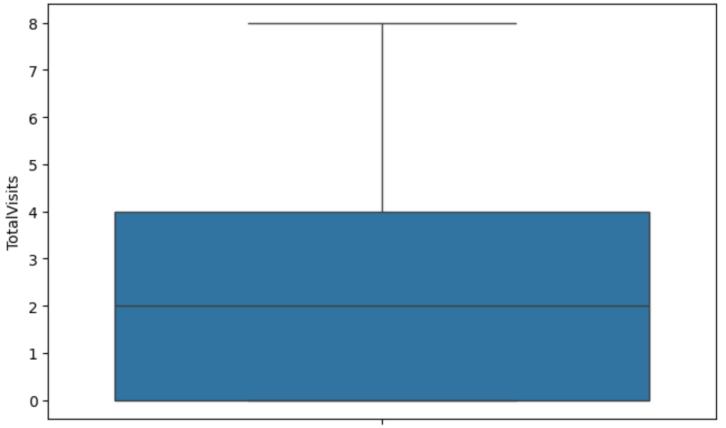
Before





After





Data Preparation

Columns with 'Yes' or 'No' values which are converted to 1 and 0 respectively

- Do Not Email
- Do Not Call
- Search
- Digital Advertisement
- Through Recommendations
- A free copy of Mastering The Interview

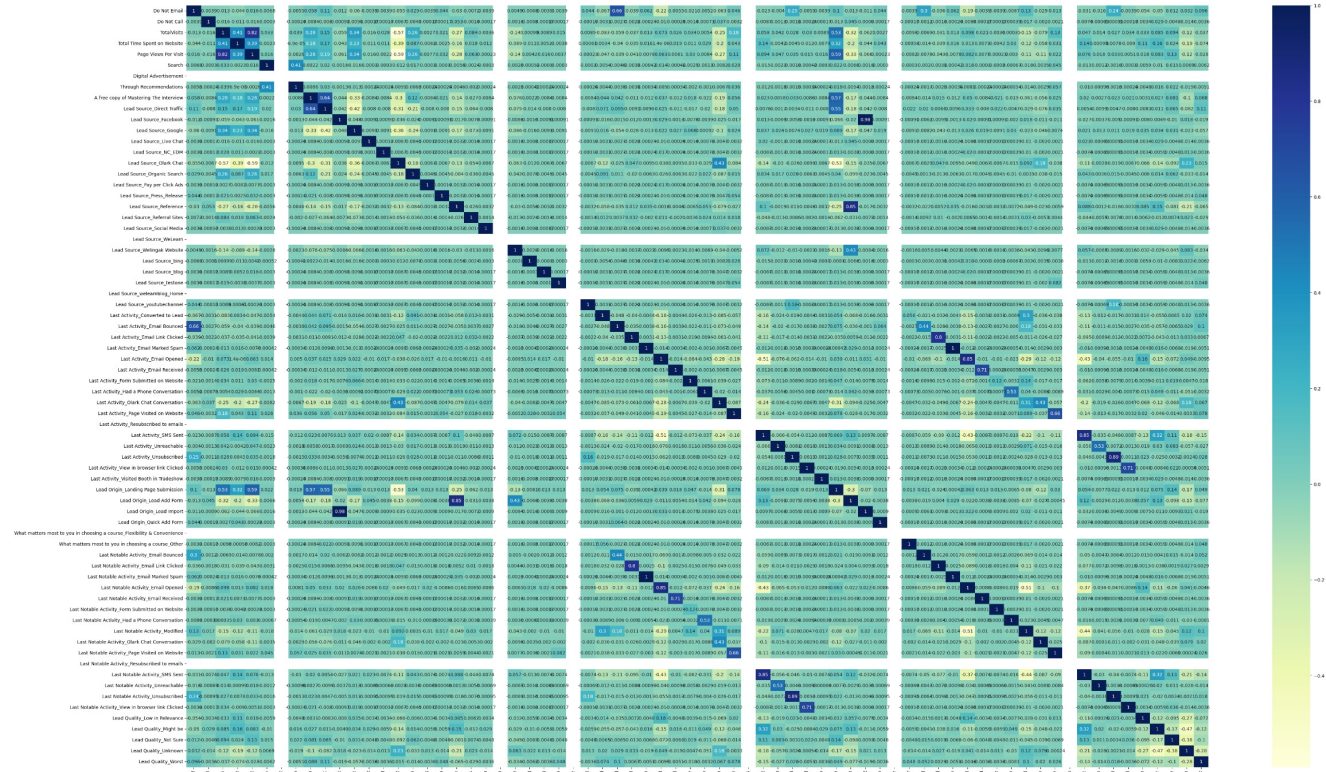
Create dummies of the categorical variables

Test-Train Split using sklearn train_test_split function

Feature Scaling using StandardScaler

Data Preparation

Checking for correlations and dropped highly correlated variables(Correlation > 0.8)



Model Building

Feature Selection using RFE

- Estimator used is LogisticRegression
- Number of features selected are 15

Using Statsmodels to build a model with the feature selected using RFE

- GLM method
- Used Binomial families
- Dropped columns with P-value(< 0.05) and VIF(< 5)

Model Building

The final model parameters are

	coef	std err	z	P>|z|	[0.025	0.975]
const	2.6151	0.166	15.758	0.000	2.290	2.940
Total Time Spent on Website	1.0628	0.041	25.776	0.000	0.982	1.144
Lead Source_Direct Traffic	-3.5659	0.181	-19.714	0.000	-3.920	-3.211
Lead Source_Google	-3.1762	0.176	-18.076	0.000	-3.521	-2.832
Lead Source_Olark Chat	-2.2344	0.178	-12.546	0.000	-2.584	-1.885
Lead Source_Organic Search	-3.3349	0.196	-16.986	0.000	-3.720	-2.950
Lead Source_Referral Sites	-3.4714	0.386	-8.991	0.000	-4.228	-2.715
Lead Source_Welingak Website	2.9979	1.023	2.930	0.003	0.993	5.003
Last Activity_Email Bounced	-2.0822	0.355	-5.863	0.000	-2.778	-1.386
Last Activity_Olark Chat Conversation	-1.2866	0.201	-6.400	0.000	-1.681	-0.893
Last Notable Activity_Modified	-0.8700	0.083	-10.478	0.000	-1.033	-0.707
Last Notable Activity_Olark Chat Conversation	-0.8347	0.374	-2.234	0.025	-1.567	-0.102
Lead Quality_Might be	1.5081	0.096	15.719	0.000	1.320	1.696
Lead Quality_Not Sure	-0.5484	0.112	-4.884	0.000	-0.768	-0.328
Lead Quality_Worst	-3.5711	0.470	-7.592	0.000	-4.493	-2.649

	Features	VIF
9	Last Notable Activity_Modified	2.04
8	Last Activity_Olark Chat Conversation	2.00
3	Lead Source_Olark Chat	1.62
2	Lead Source_Google	1.49
1	Lead Source_Direct Traffic	1.47
10	Last Notable Activity_Olark Chat Conversation	1.38
0	Total Time Spent on Website	1.28
11	Lead Quality_Might be	1.24
12	Lead Quality_Not Sure	1.22
4	Lead Source_Organic Search	1.18
13	Lead Quality_Worst	1.15
7	Last Activity_Email Bounced	1.12
5	Lead Source_Referral Sites	1.02
6	Lead Source_Welingak Website	1.01

Confusion matrix

Actual \ Predicted	No	Yes
No	2639	963
Yes	308	1898

Accuracy
(TP + TN/(TP + FP + TN + FN))
0.781164

Sensitivity
(TP / TP+FN)
0.8603807796917498

Specificity
TN / TN+FP
0.7326485285952249

Calculate false postive rate - predicting converted when customer does not have converted
FP/ TN+FP
0.2673514714047751

Positive predictive value
TP / TP+FP
0.6634044040545264

Negative predictive value
TN / TN+ FN
0.8954869358669834

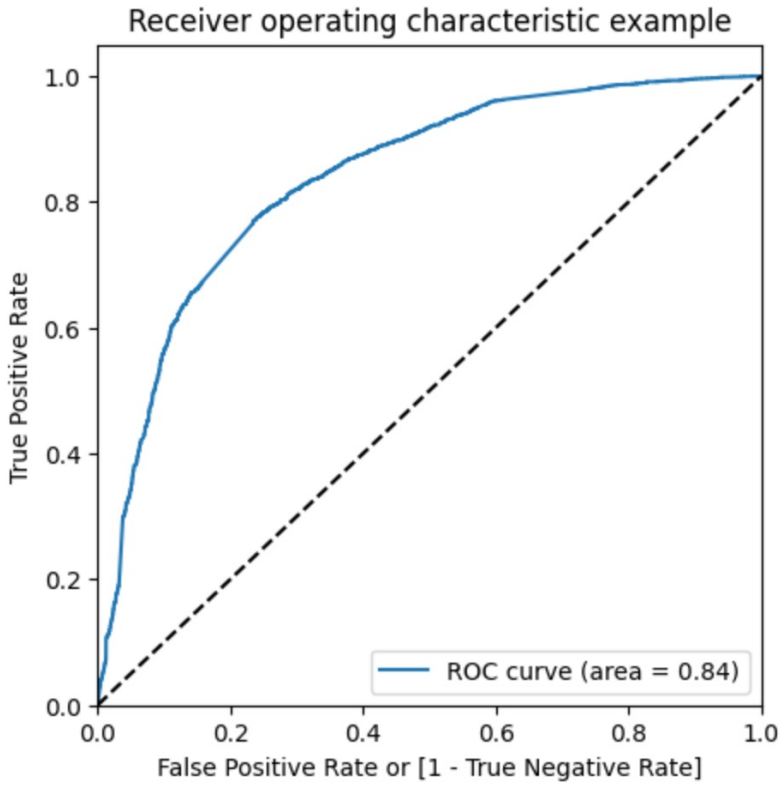
Precision
TP / (TP + FP)
0.6634044040545264

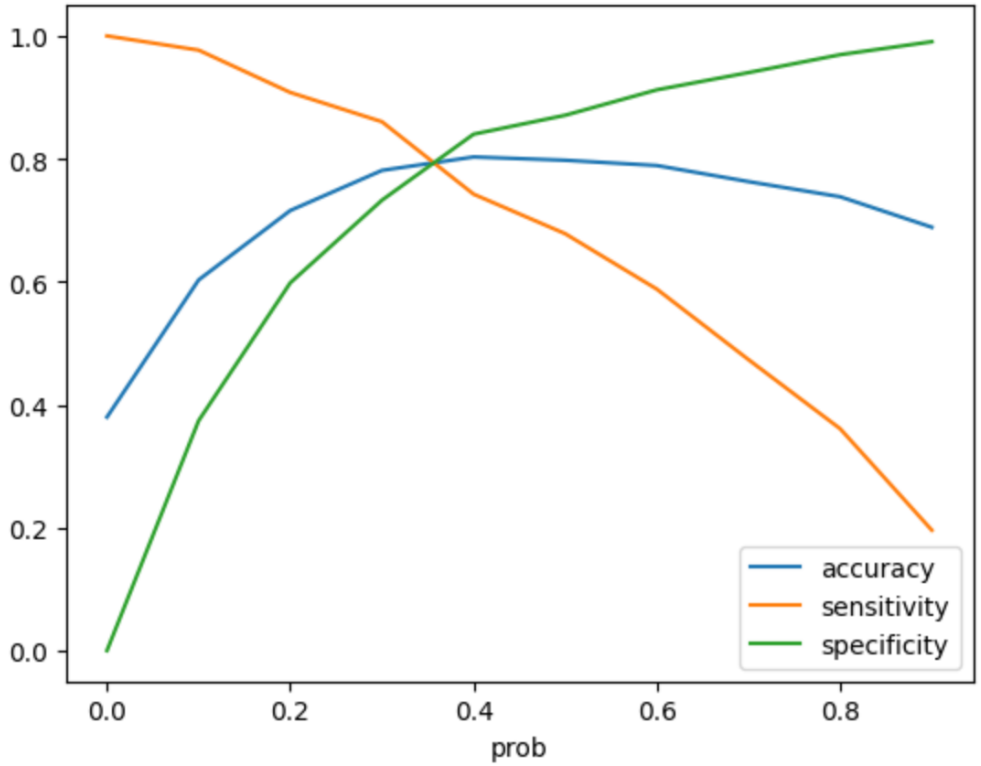
Recall
TP / (TP + FN)
0.8603807796917498

ROC Curve and Plot of various cut-offs

An ROC curve demonstrates several things:
It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity). The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

The second graph shows the accuracy, sensitivity and specificity of various cut-offs. We chose 0.3 a the cut-off because sensitivity is high in 0.3 compared to 0.4 which will help us to identify more conversions which will reduce the False Negatives. (TP / (TP+FN)). The graph also recommends an optimal cut-off between 0.3 and 0.4.





Results of various cut-offs

Below show the predictions using various cut-off and a final prediction with cut-off as 0.3

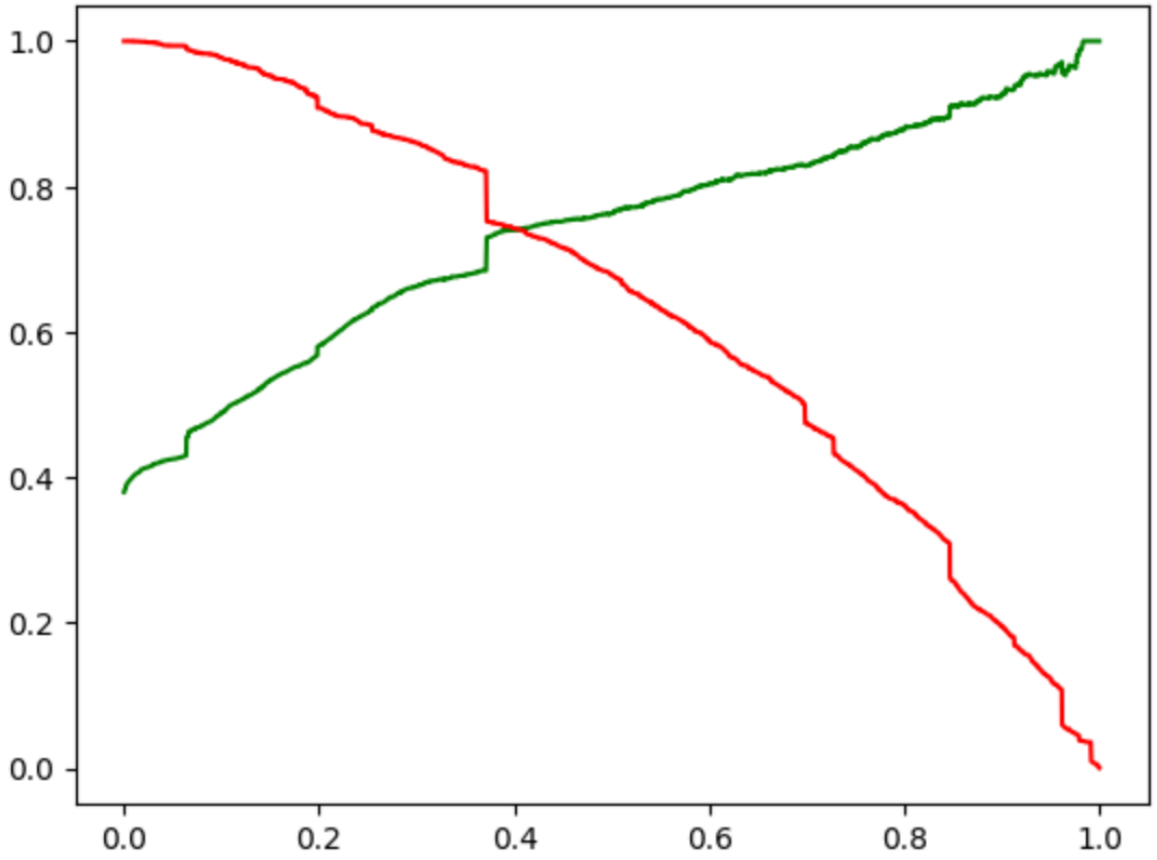
It also shows the **Lead Score** associated with each applicant which can be used to analyze hot leads

The higher the lead score the higher are the chances that the lead will get converted.

	Converted	Converted_Prob	Prosepct ID	predicted	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	final_predicted	Lead Score
0	0	0.912598	8076	1	1	1	1	1	1	1	1	1	1	1	1	91.259753
1	0	0.063961	8036	0	1	0	0	0	0	0	0	0	0	0	0	6.396121
2	0	0.001918	3744	0	1	0	0	0	0	0	0	0	0	0	0	0.191807
3	1	0.254427	5180	0	1	1	1	0	0	0	0	0	0	0	0	25.442655
4	0	0.063961	8866	0	1	0	0	0	0	0	0	0	0	0	0	6.396121

Precision and Recall tradeoff

We have selected 0.3 as the cut-off. According to the above precision recall curve a good cut-off may be between 0.3 to 0.4. For now, we are going ahead with 0.3



Train and Test set parameters

The values of the train and test data are similar, which is a positive indicator

Train set values

Accuracy	Sensitivity	Specificity
0.781164	0.860381	0.732649

Test set values

Accuracy	Sensitivity	Specificity
0.767871	0.867109	0.711405

Recommendations

Top Three Variables and its influence on the Lead conversion

1. Total Time Spent on Website

Coefficient: 1.0628

P-value: 0.000

Interpretation: A higher amount of time spent on the website significantly increases the probability of lead conversion.

2. Lead Quality_Might be

Coefficient: 1.5081

P-value: 0.000

Interpretation: If the lead quality is marked as "Might be," it significantly increases the likelihood of conversion.

3. Lead Source_Welingak Website

Coefficient: 2.9979

P-value: 0.003

Interpretation: Leads coming from the "Welingak Website" source have a strong positive influence on conversion probability.

Explanation:

Total Time Spent on Website has a positive coefficient of 1.0628, indicating that as the time spent on the website increases, the probability of conversion also increases.

Lead Quality_Might be has the a positive coefficient (1.5081) among categorical variables, making it a strong predictor.

Lead Source_Welingak Website has a very high positive coefficient (2.9979), indicating that leads from this source are more likely to convert.

Aggressive Lead conversion strategies

1. Prioritize High Probability Leads:

Focus on Hot Leads: Use the model predictions to identify leads with high conversion probabilities (e.g., Converted_Prob > 0.8). These leads should be the top priority for phone calls and follow-ups.

Intern Involvement: Assign these high-probability leads to the interns, ensuring that the most promising leads receive immediate and focused attention.

2. Segment Leads Based on Probability:

Tier 1: Leads with Converted_Prob > 0.8 - These leads should receive personal calls and follow-ups as soon as possible.

Tier 2: Leads with Converted_Prob between 0.5 and 0.8 - These leads should receive follow-up emails and calls but can be handled with slightly less urgency.

Tier 3: Leads with Converted_Prob < 0.5 - These can receive automated or less frequent follow-ups, allowing the team to focus on higher-priority leads.

3. Optimize Intern Resources:

Training and Scripts: Ensure interns are trained to handle common objections from customers and equipped with scripts that have proven effective in converting leads.

Monitoring and Feedback: Regularly monitor the interns' progress and provide feedback to optimize their interactions with leads.

4. Enhance Communication Channels:

Multichannel Approach: Use a combination of phone calls, personalized emails, and possibly SMS to engage with leads. The more touchpoints, the higher are the chances visibility and conversion.

Follow-Up Strategy: Implement a structured follow-up strategy where leads receive multiple contacts over a short period, creating urgency but not in a disturbing manner. The time slots when they receive the calls should be allocated accordingly.

5. Analyze and Adjust:

Daily Review: Each day, review the conversion data to adjust strategies and ensure that efforts are focused on the most promising leads.

Use Data Insights: Continuously analyze which strategies are working best and adjust the approach based on real-time data.

By concentrating efforts on the high-probability leads and leveraging the additional workforce provided by the interns, X Education can increase its lead conversion rate significantly during this aggressive phase.

Optimization strategy post target achievement

1. Focus on High-Probability Leads Only

2. Automate Low-Priority Communication

3. Reassign Sales Team to New Tasks

4. Engage in Relationship Building

5. Focus on Long-Term Projects

By adopting this strategy, X Education can minimize unnecessary phone calls, efficiently allocate the sales team's time, and set the environment for continued success in future quarters.

Thank You