LEADS SCORING CASE STUDY

LE TRUNG KIEN - Data Science Program
(Global) C12

Problem Statement

X Education sells online courses. They recruit students from various channels including websites, search engines.

Via marketing channels, they get a number of people who are interested in the provided courses, known as ‘leads’.

Their sales team need to convert these leads into actual students by different means (phone calls, sms, emails,...).

The current conversion rate is around 30%.

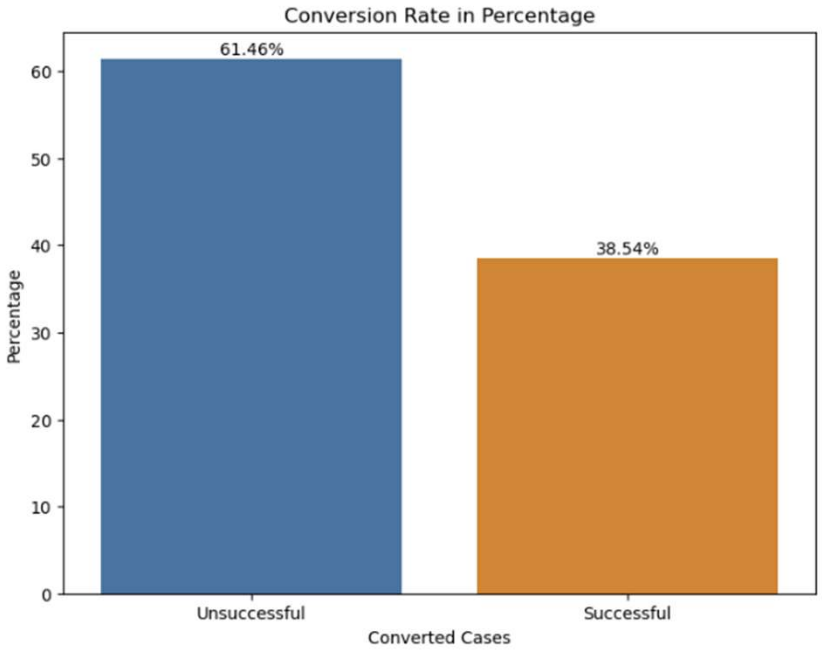
X Education wants to improve the rate to about 80%.

Solution

Data Clearning – EDA – Data Preparation – Model Building

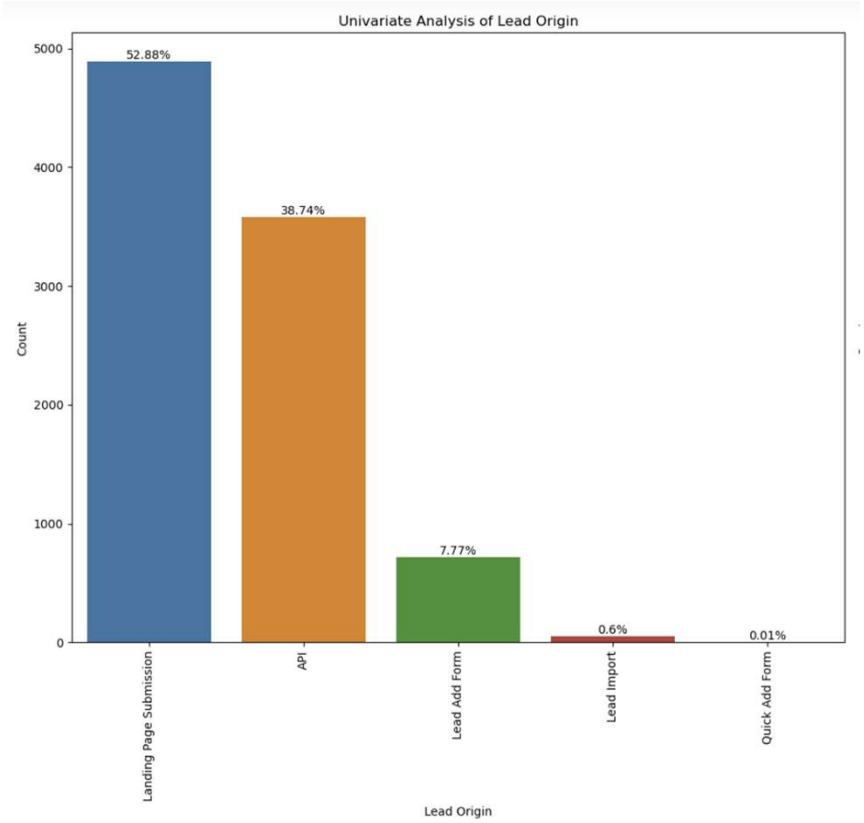
Data Cleaning

1. Drop unwanted columns
2. Deal with missing values
3. Treat outliers
4. Group low frequencies
5. Map binary categorical values
6. Standardize values



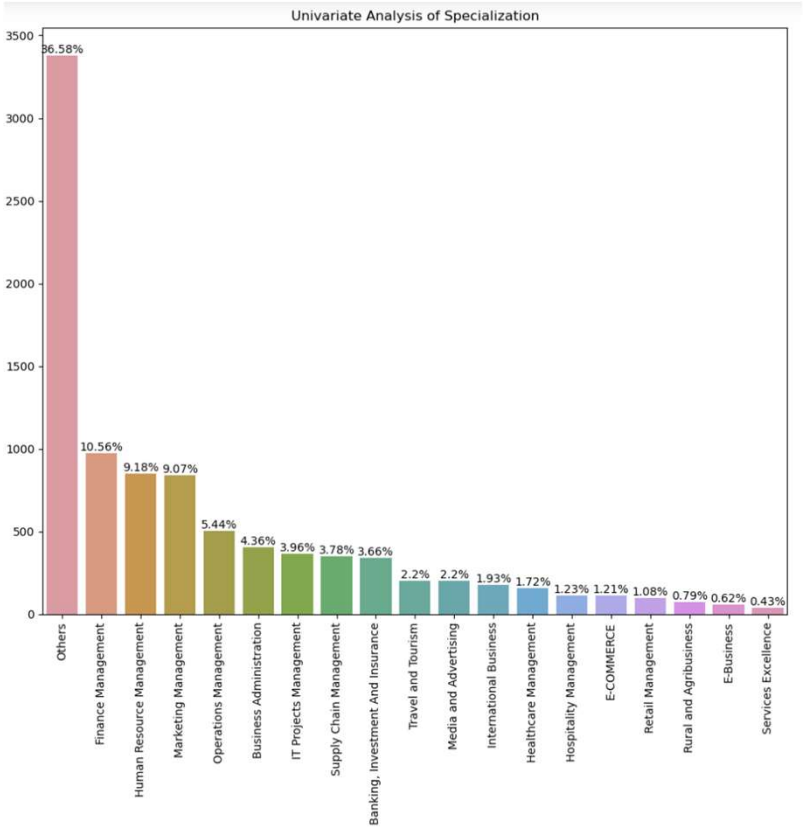
EDA

Check data imbalance: that only 38.54% of leads were successfully converted.



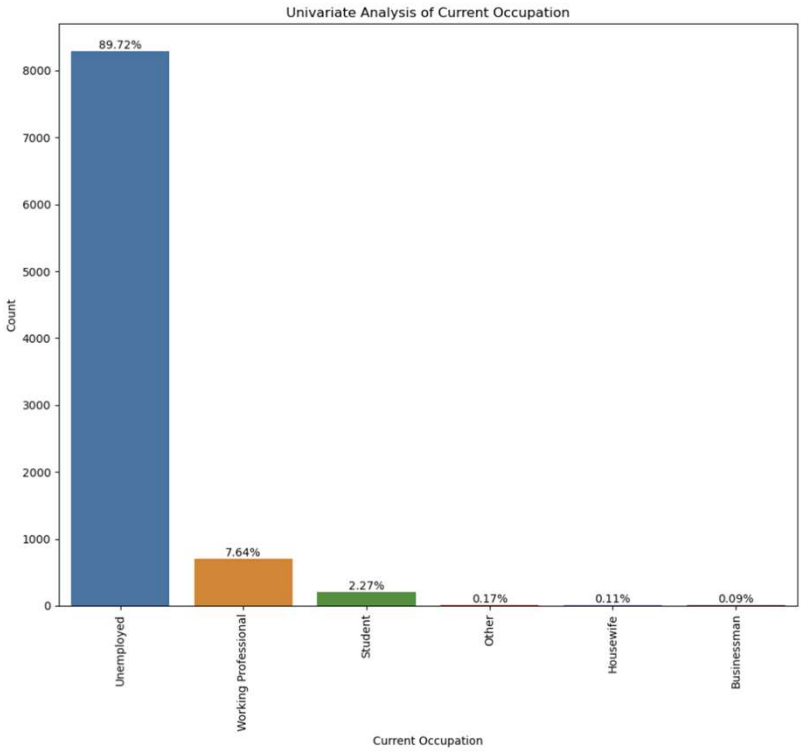
EDA: Univariate analysis

The majority of leads come from "Landing Page Submission" (52.88%) and "API" (38.74%).



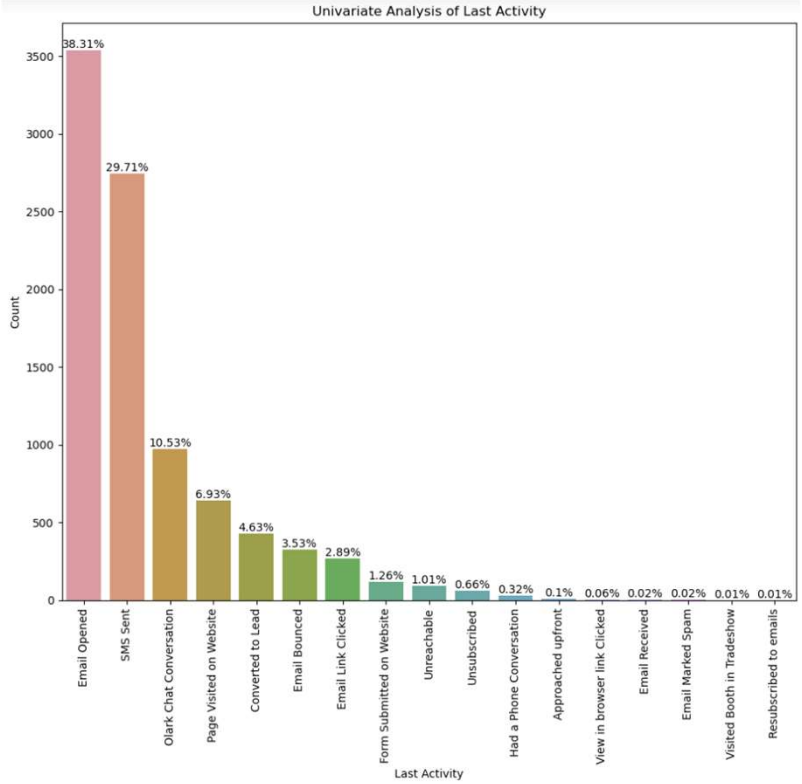
EDA: Univariate analysis

Of the known specialization, people who worked in Finance Management, HRM, and Marketing Management have a higher tendency to be attracted by the courses provided by X Education.



EDA: Univariate analysis

Nearly 90% of leads are unemployed.

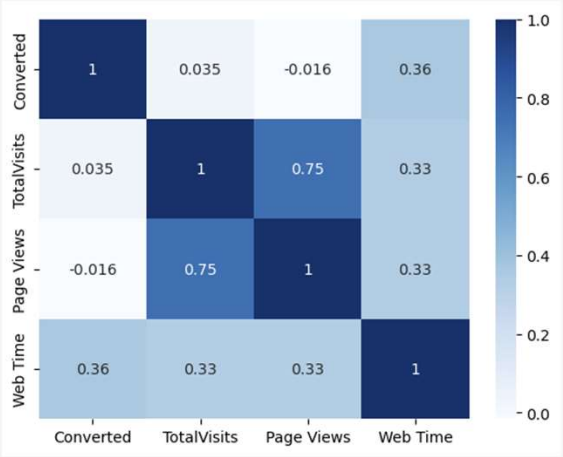


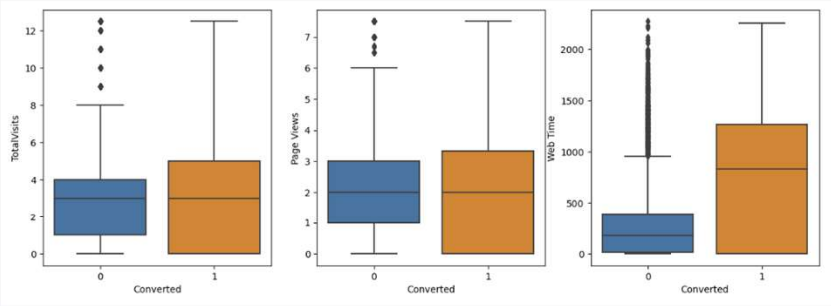
EDA: Univariate analysis

Leads seem to be more keen on opening emails and sending sms.

EDA: Bivariate analysis

- Landing Page Submission accounts for about 53% of the leads, while that of API and Lead Add Form are 29% and 8% respectively. The Conversion rates of Landing Page Submission, API, and Lead Add Form are, therefore, 31%, 36%, and 88% accordingly.
- Of all the sources, leads that come from Reference and Welingak Website are mostly sucessfully converted.
- Approximately 63% of potential customers who sent an SMS become an X student, the highest conversion rate based on Last Activity. Top lead contributors are groups who opened email (38%), sent SMS (30%), or participated in olark chat (11%). Olark Chat seems to be ineffective since only 9% of clients were converted.
- Of the known categories, Finance Management (11%), HRM (9%), and Marketing Management (9.9%) are more significant compared to other groups. These groups' conversion rate are fairly high (>44%) with Marketing Management almost reached 50% of conversion.
- Although the majority of leads come from people who do not have a job, only 34% of them participated in X's courses. Working Professionals account for just 8% of leads but this group's conversion rate is remarkably high (88%).
- Distributing a free copy of Mastering The Interview does not seem to improve conversion rate as only 35% of the people who received Mastering the Interview are converted comparing with 40% for those who did not.





EDA: Bivariate analysis

It seems that leads who spent more time on the website have a higher tendency to be converted.

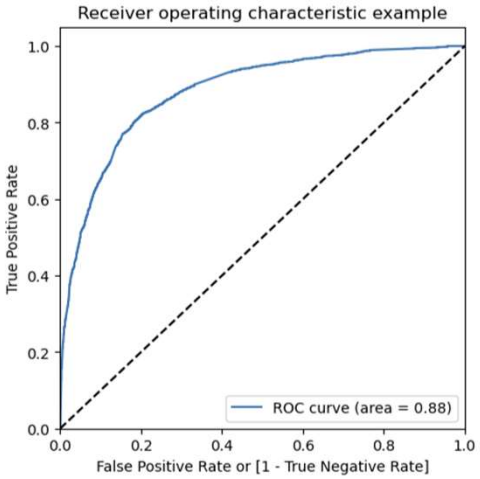
Data Preparation

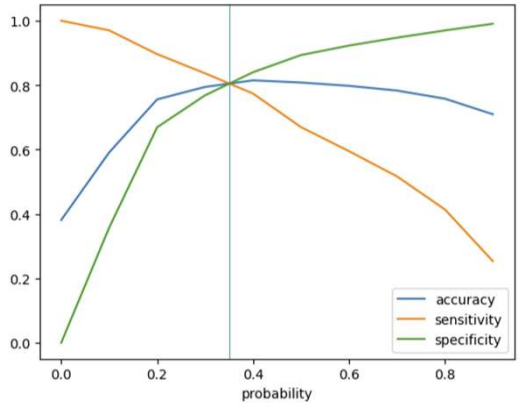
Create dummy variables for categorical variables.

Split the Train – Test sets at 70:30 ratio.

Scale numerical data with the MinMaxScaler

Model Building





Area under ROC curve: 0.88/1

Optimal cutoff point: 0.35

Model Building

Model 1: RFE to select 15 most influential features

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6452
Model Family:	Binomial	Df Model:	15
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2703.8
Date:	Tue, 31 Oct 2023	Deviance:	5407.5
Time:	21:22:38	Pearson chi2:	9.91e+03
No. Iterations:	21	Pseudo R-squ. (CS):	0.3894
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025
const	-1.3708	0.154	-8.897	0.000	-1.673
TotalVisits	1.1413	0.202	5.646	0.000	0.745
Current Occupation_Housewife	23.0073	1.34e+04	0.002	0.999	-2.62e+04
Specialization_Others	-1.0996	0.122	-9.000	0.000	-1.339
Last Activity_Olark Chat Conversation	-1.3646	0.169	-8.077	0.000	-1.696
Last Activity_Had a Phone Conversation	2.0626	0.666	3.097	0.002	0.757
Last Activity_Email Bounced	-1.8248	0.296	-6.164	0.000	-2.405
Last Activity_Converted to Lead	-1.0752	0.222	-4.841	0.000	-1.510
Lead Source_Welingak Website	2.4996	0.750	3.332	0.001	1.029
Lead Source_Olark Chat	1.0286	0.136	7.571	0.000	0.762
Last Activity_SMS Sent	1.2756	0.075	17.113	0.000	1.130
Current Occupation_Working Professional	2.7246	0.191	14.287	0.000	2.351
Web Time	4.4355	0.162	27.366	0.000	4.118
Page Views	-0.9852	0.217	-4.530	0.000	-1.411
Lead Origin_Lead Add Form	3.0167	0.212	14.211	0.000	2.601
Lead Origin_Landing Page Submission	-1.0854	0.127	-8.532	0.000	-1.335

Model Building

Model 2: remove variables with p>0.05

Generalized Linear Model Regression Results

ep. Variable:	Converted	No. Observations:	6468
odel:	GLM	Df Residuals:	6453
odel Family:	Binomial	Df Model:	14
ink Function:	Logit	Scale:	1.0000
ethod:	IRLS	Log-Likelihood:	-2711.2
ate:	Tue, 31 Oct 2023	Deviance:	5422.5
ime:	21:24:10	Pearson chi2:	9.94e+03
o. Iterations:	7	Pseudo R-squ. (CS):	0.3880
ovariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025
onst	-1.3552	0.154	-8.808	0.000	-1.657
otalVisits	1.1244	0.202	5.570	0.000	0.729
pecialization_Others	-1.1046	0.122	-9.046	0.000	-1.344
ast Activity_Olark Chat Conversation	-1.3687	0.169	-8.103	0.000	-1.700
ast Activity_Had a Phone Conversation	2.0524	0.665	3.084	0.002	0.748
ast Activity_Email Bounced	-1.8366	0.296	-6.204	0.000	-2.417
ast Activity_Converted to Lead	-1.0873	0.222	-4.899	0.000	-1.522
ead Source_Welingak Website	2.4853	0.750	3.313	0.001	1.015
ead Source_Olark Chat	1.0216	0.136	7.528	0.000	0.756
ast Activity_SMS Sent	1.2659	0.074	17.008	0.000	1.120
urrent Occupation_Working Professional	2.7166	0.191	14.250	0.000	2.343
eb Time	4.4322	0.162	27.380	0.000	4.115
age Views	-0.9892	0.217	-4.553	0.000	-1.415
ead Origin_Lead Add Form	3.0248	0.212	14.267	0.000	2.609
ead Origin_Landing Page Submission	-1.0798	0.127	-8.495	0.000	-1.329

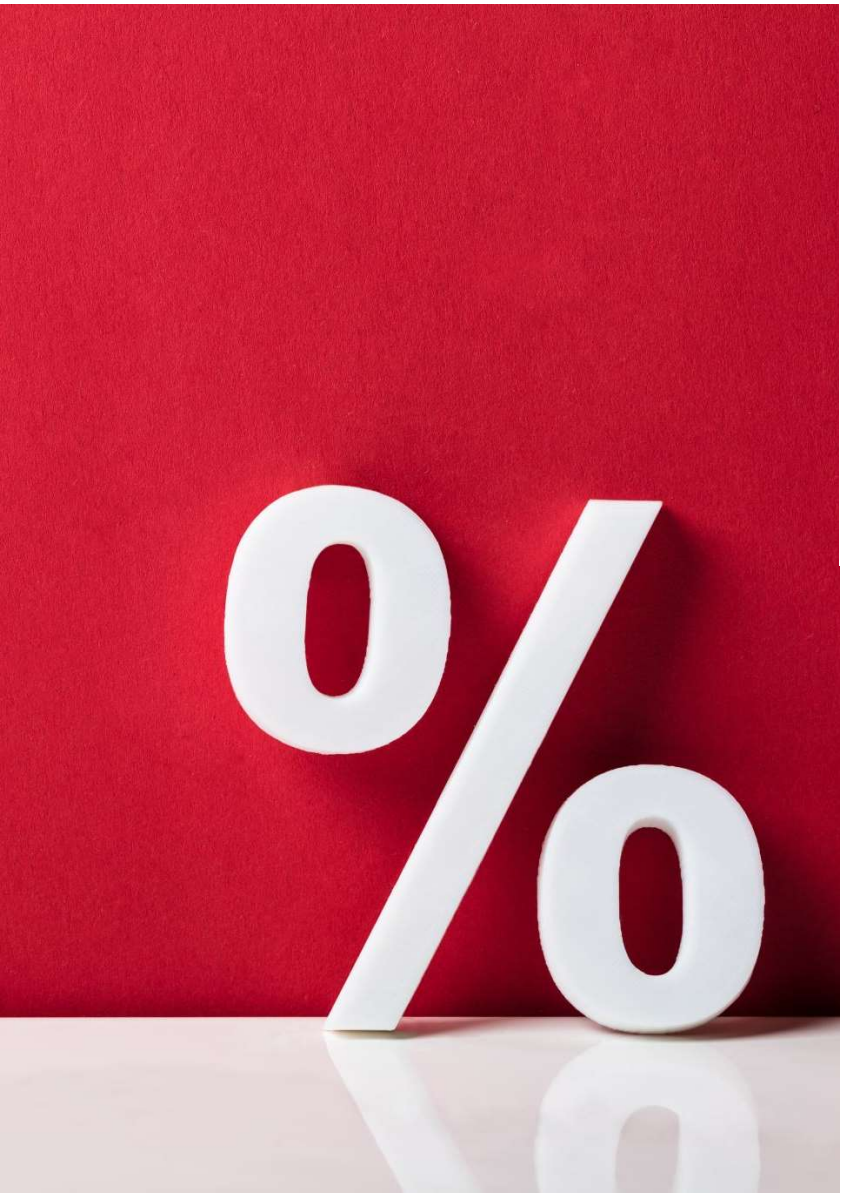
Model Building

Model 3: remove variables with VIF>5

Generalized Linear Model Regression Results

ep. Variable:	Converted	No. Observations:	6468
odel:	GLM	Df Residuals:	6454
odel Family:	Binomial	Df Model:	13
ink Function:	Logit	Scale:	1.0000
ethod:	IRLS	Log-Likelihood:	-2721.7
ate:	Tue, 31 Oct 2023	Deviance:	5443.5
ime:	21:31:39	Pearson chi2:	9.82e+03
o. Iterations:	7	Pseudo R-squ. (CS):	0.3860
ovariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.
onst	-1.5505	0.148	-10.473	0.000	-1.
otalVisits	0.6484	0.174	3.730	0.000	0.
pecialization_Others	-1.0735	0.122	-8.832	0.000	-1.
ast Activity_Olark Chat Conversation	-1.3647	0.169	-8.085	0.000	-1.
ast Activity_Had a Phone Conversation	2.0347	0.659	3.087	0.002	0.
ast Activity_Email Bounced	-1.7902	0.295	-6.071	0.000	-2.
ast Activity_Converted to Lead	-1.0267	0.222	-4.635	0.000	-1.
ead Source_Welingak Website	2.4715	0.749	3.298	0.001	1.
ead Source_Olark Chat	1.1915	0.131	9.111	0.000	0.
ast Activity_SMS Sent	1.2342	0.074	16.720	0.000	1.
urrent Occupation_Working Professional	2.7089	0.191	14.205	0.000	2.
eb Time	4.4132	0.161	27.336	0.000	4.
ead Origin_Lead Add Form	3.2095	0.208	15.418	0.000	2.
ead Origin_Landing Page Submission	-1.1317	0.126	-8.955	0.000	-1.



Conclusion

Model results:

- Accuracy 81%
- Sensitivity 75%
- Specificity 85%
- Precision 76%
- Recall 75%

Conclusion

The top 3 features that contribute to better prediciting hot leads are:

- Web Time
- Lead Origine_Lead Add Form
- Current Occupation_Working Professional

Recommendations

X Education may want to invest in upgrading their website with new features, more interesting information as those who are attracted to the website have a higher tendency to become students.

The company may also want to have a marketing strategy focusing on working professionals.

Increasing budgets for Welingak Website, Olark Chat may help boosting hot leads.

Spend more on telephone-marketing and sms campaigns.

The landing page submission as well as email bounced should be improved.

Thank you for your attention!

THE END