Lead Scoring

Submitted by
Imran Khan
Richa Dhir

Problem Statement

XEducation needs a way to help them filter most promising leads that will eventually buy products and services from them.

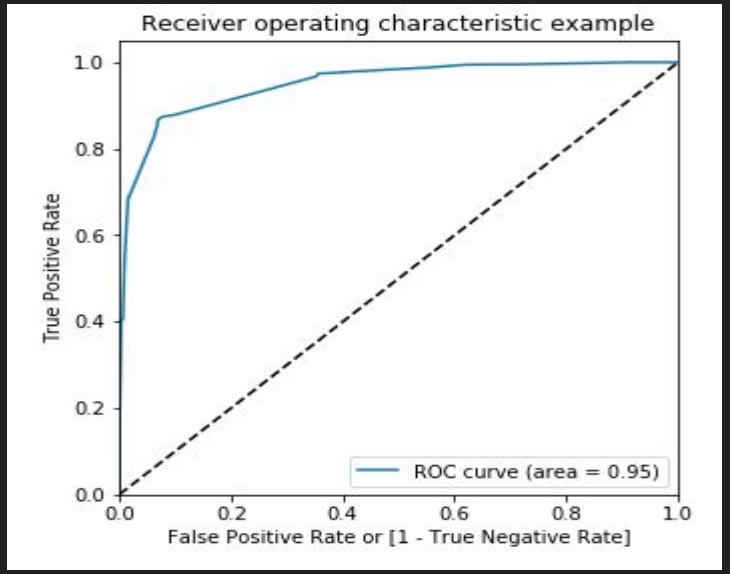
At present even though the company gets a huge bulk of leads from various sources their conversion rate is just 30%. XEducation wants to identify Hot leads/Potential leads in order to push their conversion rate beyond 80%.

Approach

- Need to devise logistic regression model that would output probability of the respective lead to get converted into a potential lead which would eventually become a paying customer for XEducation.
- Perform Univariate analysis on each column with respect to the target feature(Converted).
- Drop irrelevant features/columns that doesn’t provide any conclusive information.
- Treat Outliers, Impute missing values and create derived attributes wherever necessary to aid in the analysis process.
- Perform RFE on the final dataset to fetch the top 15 most significant features.
- Perform stepwise elimination process on based on P value to get the final list of variables that will be used for predicting the probability of lead conversions.
- Check VIF
- Based on the business requirement choose the appropriate probability cutoff to calculate lead score.

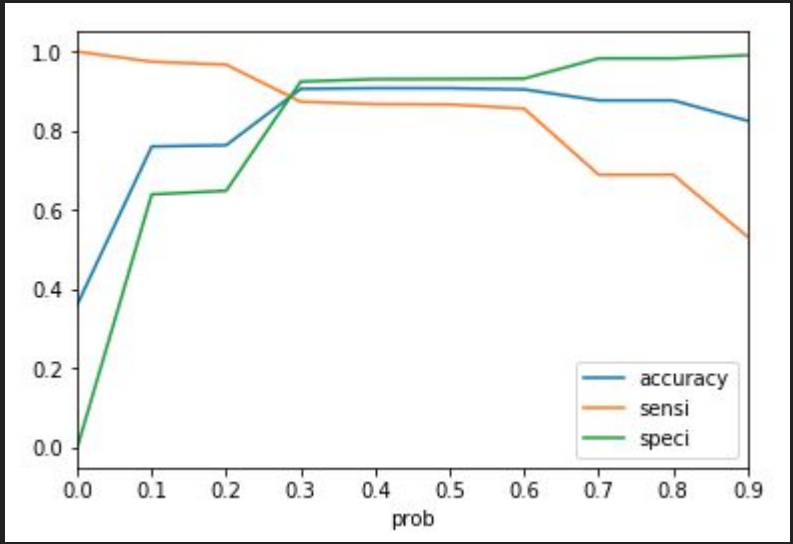
Logistic Regression - ROC

- We see True Positive and False Positive rate plotted in the below ROC curve where area under the curve is around 95% which is an indication of a good classification model.



Probability Cutoff

- From the below plot we can see very clearly the optimal probability cutoff is around 0.3 which is the best of accuracy, sensitivity and specificity achieved on the train set.
- Based on this cutoff the final accuracy achieved is around 90% on the test set.



Model Evaluation Metrics

- Sensitivity: Probability of a lead being an Actual Lead -- 87%
- Specificity: Probability of a Non lead being an Actual Non Lead -- 92%
- Positive Predictive Value(Precision): Probability that the lead was converted if it was identified as a Potential Lead -- 88%
- Negative Predictive Value(Recall): Probability that the patients was not converted if it was identified as Potential Lead -- 92%
- F1 Score -- 87%

Inferences

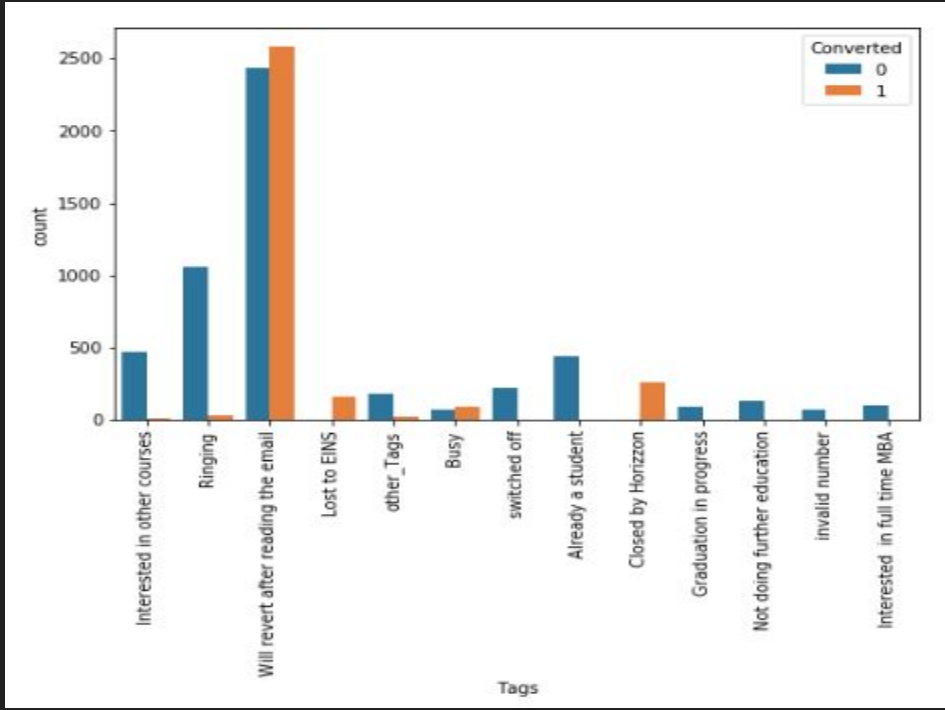
- These are the final list of features after RFE.
- Dummy variables “Tags_Closed by Horizon”, “Tags_Lost to EINS” and “Tags_Will revert after reading the email” are the most significant variables.
- Top 3 primary variables that are most significant are Tags, Lead Source and Lead Quality.

Generalized Linear Model Regression Results

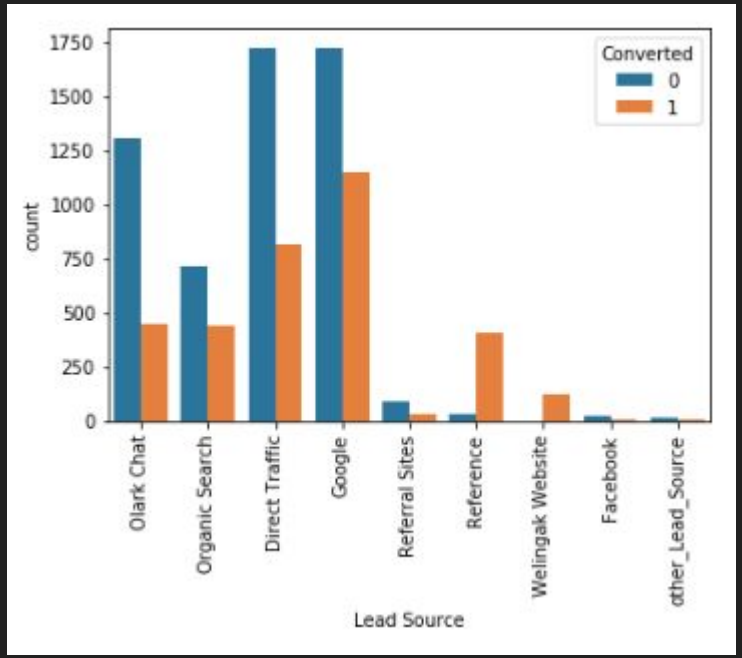
Dep. Variable:	Converted	**No. Observations:**	5618
Model:	GLM	**Df Residuals:**	5604
Model Family:	Binomial	**Df Model:**	13
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1443.5
Date:	Sat, 08 Jun 2019	**Deviance:**	2887.0
Time:	15:29:57	**Pearson chi2:**	2.18e+04
No. Iterations:	8	**Covariance Type:**	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.8945	0.194	-14.920	0.000	-3.275	-2.514
Do Not Email	-1.2776	0.219	-5.832	0.000	-1.707	-0.848
Lead Origin_Lead Add Form	1.0883	0.378	2.882	0.004	0.348	1.828
Lead Source_Welingak Website	2.7901	0.820	3.404	0.001	1.183	4.397
Last Activity_Olark Chat Conversation	-1.5397	0.214	-7.184	0.000	-1.960	-1.120
Last Activity_SMS Sent	1.9048	0.104	18.237	0.000	1.700	2.110
Occupation_Working Professional	1.4182	0.296	4.783	0.000	0.837	1.999
Tags_Busy	4.4945	0.296	15.198	0.000	3.915	5.074
Tags_Closed by Horizzon	9.1030	1.052	8.656	0.000	7.042	11.164
Tags_Lost to EINS	9.2307	0.636	14.525	0.000	7.985	10.476
Tags_Will revert after reading the email	4.6069	0.208	22.110	0.000	4.199	5.015
Tags_switched off	-1.4187	0.754	-1.883	0.060	-2.896	0.058
Lead Quality_Not Sure	-3.0918	0.134	-23.014	0.000	-3.355	-2.828
Lead Quality_Worst	-3.0074	0.907	-3.315	0.001	-4.785	-1.229

- From analysing Tags with respect to Converted target variable, we can clearly see that all the 3 attributes “Tags_Closed by Horizon”, “Tags_Lost to EINS” and “Tags_Will revert after reading the email” are very significant.



- From analysing Lead Source with respect to Converted target variable, we can observe attributes “Welingak website” and “Reference” are very significant.



Summary

- Based on various analysis performed on the data, we concluded that the best cutoff for determining the lead probability would be 0.3 and hence lead scores has been calculated based on this cutoff.
- Model has an accuracy of 90%, sensitivity of 87% and specificity of 92% which aligns with the business requirements and can help XEducation optimize their lead conversion rate.
- Based on the conversion probability leads have been assigned a score which can be used effectively for lead conversion.

Recommendations

- Sales team should be calling leads in decreasing order of lead scores assigned to each leads. This way we can ensure that most potential leads are being approached first.
- Teams should also be working on getting more leads from the sources that has high chances of conversions or has high Lead scores.
- Sales team should work on devising attractive referral policies that would lure existing customers to refer more candidates/leads. We have seen high conversion rate for this category.
- “Welingak website” is another source that sales team should work on to increase the count.
- We also observed that amount of time a lead spends on the website they have higher chances of conversion. Sales team should contact product manager to have them look at this aspect so that they can put in more effort on UX testing and to make overall platform experience more engaging.
- Sales team should be sending more promotional mails and messages to working professionals as these are the potential leads who get converted most.
- Majority of the leads come from Mumbai, hence Sales team should try to promote more aggressively in other cities as well to spread awareness.