LEAD SCORING- GROUP ASSIGNMENT

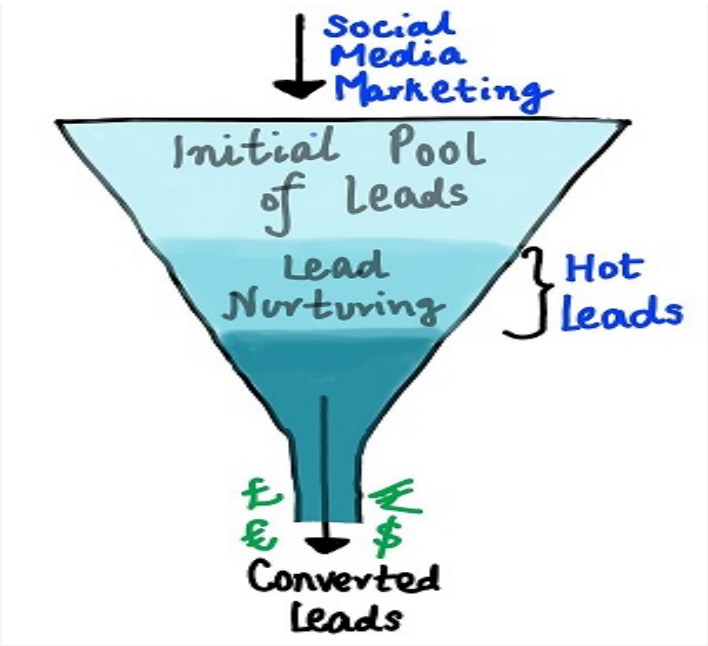
Abstract

Detailed Presentation for the assignment

Amulya Thumuluru & Eric Noel Pereira

Problem Statement: 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%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone. A typical lead conversion process can be represented using the following funnel:



Lead Conversion Process - Demonstrated as a funnel

As you can see, there are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

The goal of the case study is to 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.

Solution Summary:

The aim of the exercise is to be able to build a model which the company – X Education can utilize to convert their probable leads to customers of the company.

Data Cleaning & EDA: We started off by analysing the quality of data and doing the exploratory analysis. Quite a lot of the data points had values of 'Select' which were equivalent of null values as the users had not given any response for those data points. Hence we replaced all such data points will null values.

Following strategies were employed to clean the data and select the relevant columns:

- Columns were dropped which had:
  - More than 50% missing values: Below are the set of columns which were dropped. Figure shows the % of missing values

```
How did you hear about X Education              78.46
Lead Profile                                    74.19
Lead Quality                                    51.59
```

  - Tags Column was dropped because the main intent is to Identify the probable leads to whom the sales teams can make calls to. The assumption to drop this column was that the tags were updated against the leads after calling up the customers.
  - Data was imbalanced and skewness was detected i.e. if more than 95% of the responses were same. Following columns were dropped

Column Name	% Of one type of Values
Do Not Call	99.98
Search	99.85
Magazine	100.00
Newspaper Article	99.98
X Education Forms	99.99
Newspaper	99.98
Digital Advertisement	99.96
Through Recommendations	99.92
Receive More Updates About Our Courses	100.00
Update me on Supply Chain Content	100.00

Get updates on DM Content	100.00
I agree to pay the amount through cheque	100.00
Country	96.89

- Handling of missing data:
  - Missing data was imputed with 'Mode' statistic of the data column. This was done for the City, Current Occupation Columns.
  - Null values were replaced with 'Not Specified' These columns were dropped later during dummy column creation. Few of the columns which were treated this way were: Lead Origin, 'What is current Occupation', Specialization, Lead Source, Last Activity, 'What matters most to you in choosing a course'
- Data Aggregation

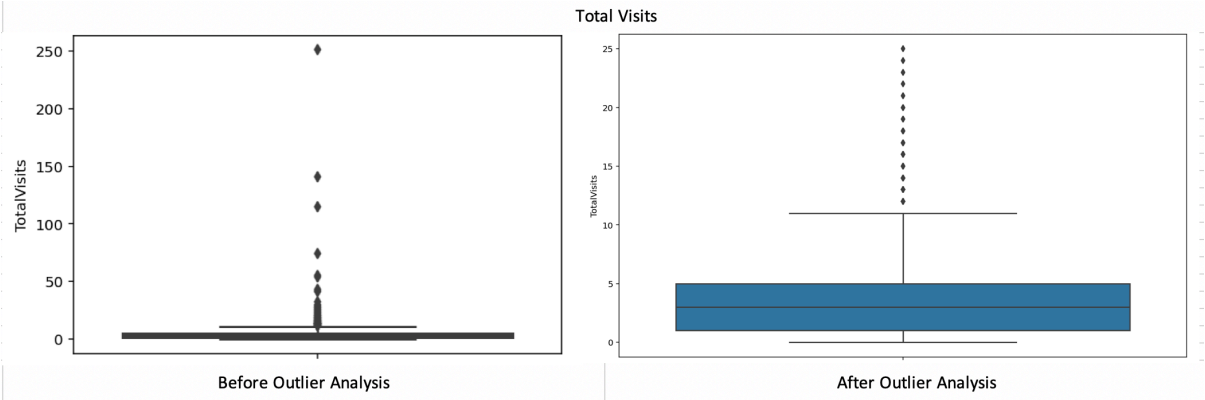
  A lot of features had categories of data which were provided as responses by a very small set of customers. For such kind of categories, we aggregated them to form a consolidated category . Few data aggregations which we did were:

  - Lead Sources: Sources with less counts were aggregated under 'Other' category
  - Last Activity: Sources with less counts were aggregated under 'Other' category
  - Specializations: Management related specializations were tagged under 'Management Specialization' category.

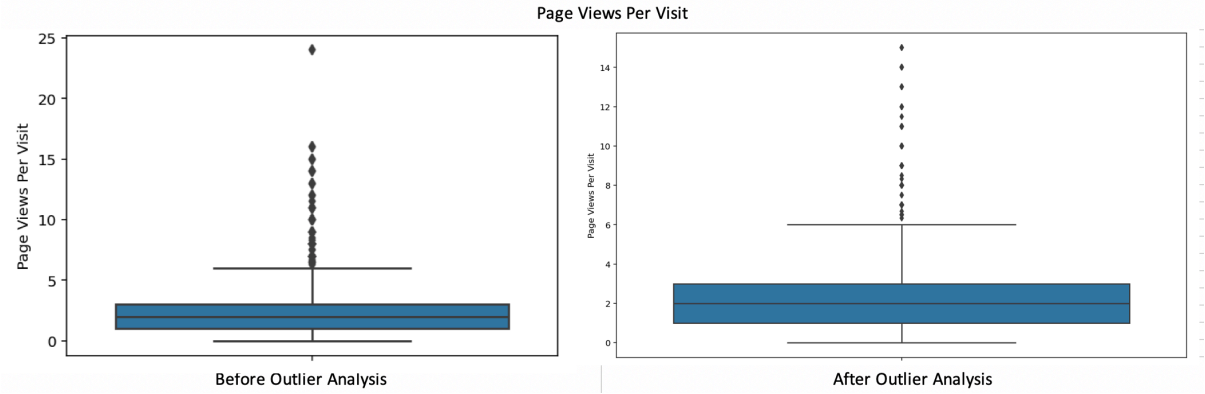
- Outlier Analysis:

  Outliers were removed from columns using the standard approach of removing dta above and below a certain threshold (Q1-1.5*IQR<X<Q3+1.5*IQR) Where Q1,Q3 were the first and the third quartiles and IQR was the Inter Quartile Range. Few columns where outlier Treatment was done was :

  Total Visits:



Page Views per Visit:



- Dummy value creation

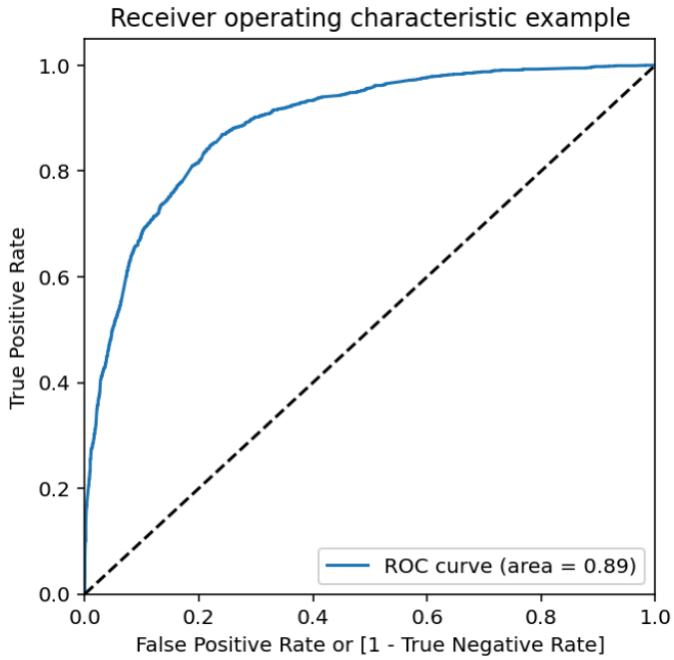
  All categorical column data was converted to dummy columns. Where-ever the null values had been imputed with 'Not Specified' those dummy columns were dropped.

Modelling: Once the data was cleaned and ready for modelling, we started off by scaling the data so that all features were on the same scale. We started to build a Logistic Regression model by first narrowing down on the total number of features using RFE to a total of 15 features. Once the initial model was built , we set the initial cut-off level at 0.5 probability score and obtained the predicted probabilities of our trained dataset. At this level we obtained the accuracy score=0.817.

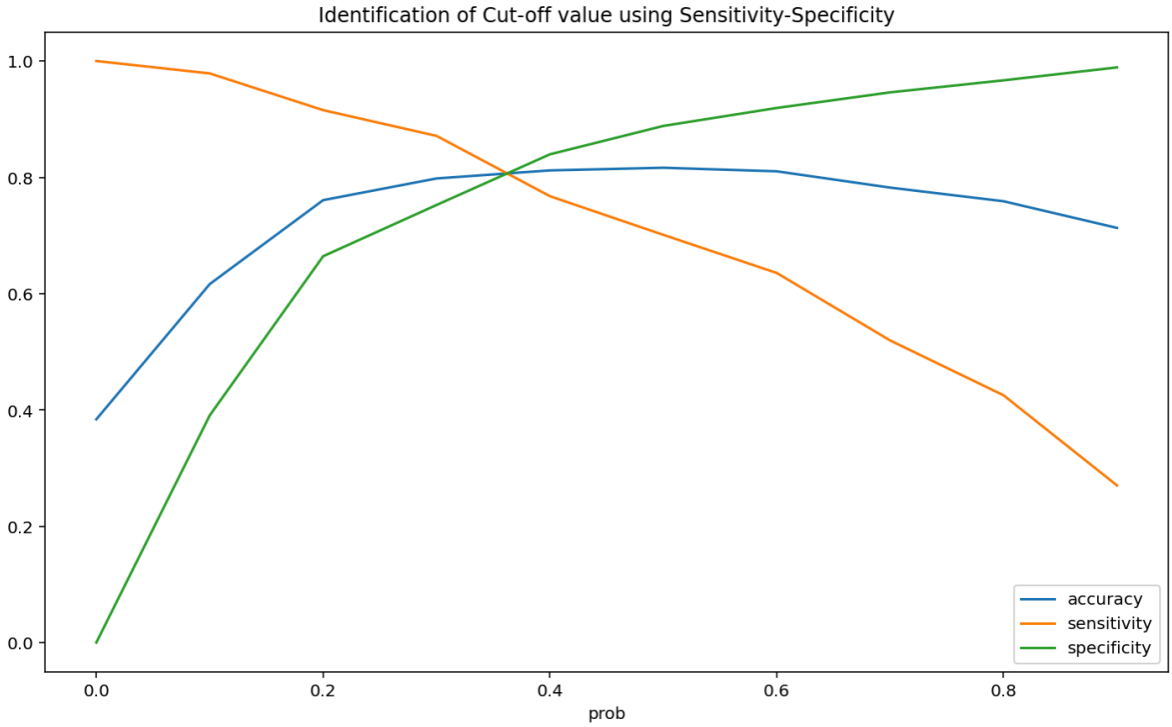
Next we checked the VIF score to remove cases of high multicollinearity but found that all features were within acceptable ranges. Next we proceeded to drop the features with high p-values and iteratively refined the model to finally arrive at a model where all p-values were significant. The final model still maintained the accuracy score of 0.817. Below were the metrics of the model:

Dep. Variable:	Converted	**No. Observations:**	6353
Model:	GLM	**Df Residuals:**	6338
Model Family:	Binomial	**Df Model:**	14
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2579.6
Date:	Sun, 19 Apr 2020	**Deviance:**	5159.2
Time:	20:13:37	**Pearson chi2:**	6.59e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

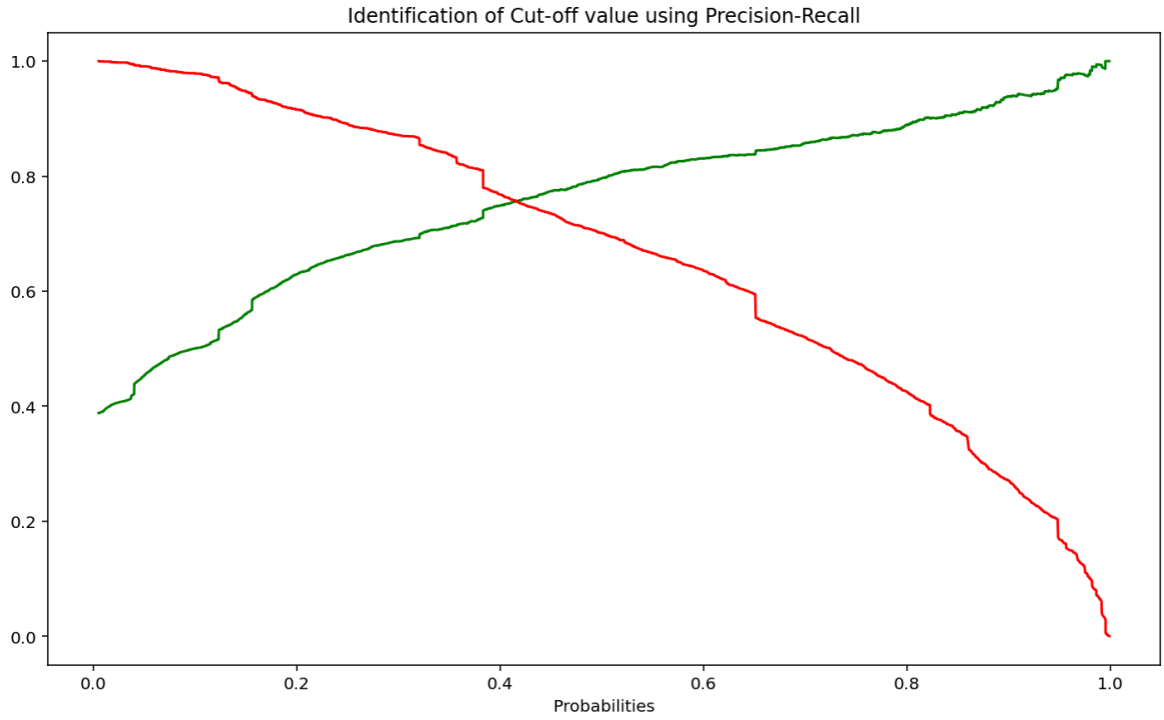
We verified the model performance by generating the ROC curve where we got a AUC =0.89 which validated our model performance as good.



To further improve the model , we plotted Sensitivity, Specificity, and Accuracy to arrive at a cut-off value of 0.3



We also checked the Precision and Recall metrics and obtained the cut-off to be 0.425.



Model performance was higher at 0.3 and hence we selected the Sensitivity-Specificity cut-off as the ideal cut-off value. Using this the final model was run on the test data. Following were the results:

Sensitivity: 0.87 Specificity:0.75 Accuracy Score:0.79

Dep. Variable:	Converted	**No. Observations:**	6353
Model:	GLM	**Df Residuals:**	6338
Model Family:	Binomial	**Df Model:**	14
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2579.6
Date:	Sun, 19 Apr 2020	**Deviance:**	5159.2
Time:	20:40:43	**Pearson chi2:**	6.59e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

Generalized Linear Model Regression Results

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.9854	0.099	-9.994	0.000	-1.179	-0.792
Total Time Spent on Website	1.1124	0.041	27.435	0.000	1.033	1.192
Lead Origin_Lead Add Form	3.6650	0.231	15.868	0.000	3.212	4.118
CurrOccupation_Working Professional	2.4688	0.187	13.179	0.000	2.102	2.836
MostMattersInChoosing_Better Career Prospects	1.2117	0.088	13.810	0.000	1.040	1.384
Lead Source_Olark Chat	1.3791	0.105	13.124	0.000	1.173	1.585
Lead Source_Welingak Website	1.8469	0.759	2.434	0.015	0.360	3.334
Last Activity_Email Bounced	-2.6555	0.391	-6.792	0.000	-3.422	-1.889
Last Activity_Olark Chat Conversation	-1.2100	0.189	-6.418	0.000	-1.580	-0.840
Last Activity_Page Visited on Website	-1.2858	0.157	-8.170	0.000	-1.594	-0.977
Last Notable Activity_Email Link Clicked	-1.9340	0.290	-6.677	0.000	-2.502	-1.366
Last Notable Activity_Email Opened	-1.0988	0.085	-12.876	0.000	-1.266	-0.932
Last Notable Activity_Modified	-1.3740	0.095	-14.468	0.000	-1.560	-1.188
Last Notable Activity_Olark Chat Conversation	-1.4557	0.374	-3.894	0.000	-2.188	-0.723
ActivityIdx_03.Low	-2.0603	0.302	-6.812	0.000	-2.653	-1.468

Inferences & Conclusion: Basis the final model, the most significant features which will aid in converting a lead are the ones with a positive co-effecient. We find the following:

Top Significant Features are:

1. Lead Origin
2. Lead Source
3. Current Occupation
4. Response of the person for the question " Whats matters most in Choosing a Course"

Under these features, the following combination of the responses yielded the most positive outcomes i.e. a person with these responses was most likely to get converted:

- For the column Lead Origin – The leads under “Lead Add Form” had high conversion i.e. Out of 718 customers who gave these responses, approx. 92% of them converted.
- For the Column Lead Source – if the source was “Welingak Website”, the conversion ratio was approx. 99%(142 responses) and if the response was “Olark Chat”, the conversion rate was approx. 75%(1755 resonses)
- For the question “What is your Current Occupation” – The responses with “Working Professional” had conversion rate of approx. 92%(706 responses).

- For the question " What matters most to you in choosing a course " - response of "Better Career prospects" – Approx 50% Conversion Ratio

Few Insights:

- Those that customers were who visited the Welingak site or interacted using the Olark Chat had quite high conversion ratios, hence such people need to be targeted first.
- As we can see that the people are interested more on Management and its allied courses and most of the people who were converted into leads are either unemployed or working professionals who opt for better career prospects , hence we can focus on such category of people more as the conversion rates were higher. Hence the interns can be asked to prioritize calling such people.
- Finally 1 in 2 customers who mentioned the reason for choosing the course as “Better Career Prospects” finally converted and it will make sense to employ the interns to target these category of people.

If the company completes its targets in a quarter, the company personnel can focus on the below activities :

- Working on data collection for fields that had “Select” as the option so that we can get more correct and accurate data
- The percentage of follow ups for those categories of customers that was mentioned as “Do Not Email – No “is [Leads -1 : 40.47% and No leads – 0 : 59.53%] so the quality follow ups can be made to those customers
- Increase of Lead source cases where more advertisements can be posted in lead sources like Wellingak Website, Olark Chat and Direct Traffic where the scope of getting the leads converted is more.