upGrad

LEAD SCORE CASE STUDY USING LOGISTIC REGRESSION

Date – 20th June 2023

Batch no – DS – C51

Submitted by:

- Manit Aslot
- Rashmi S
- Sanchita Sharma

upGrad

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%.

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Objective of this 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.

MODEL APPROACH

Reading and Understanding the data	Importing the data	Understanding the data dictionary and data set.
Data Cleaning	Handling missing values	Outlier analysis
Exploratory Data Analysis	Univariate analysis	Bivariate analysis
Data Preparation	Creating dummy variables	Train and Test data split
Model Building	Feature selection using RFE	p-value and VIF check
Model Evaluation	Evaluation of metrics such as sensitivity, specificity, accuracy	Inference

upGrad

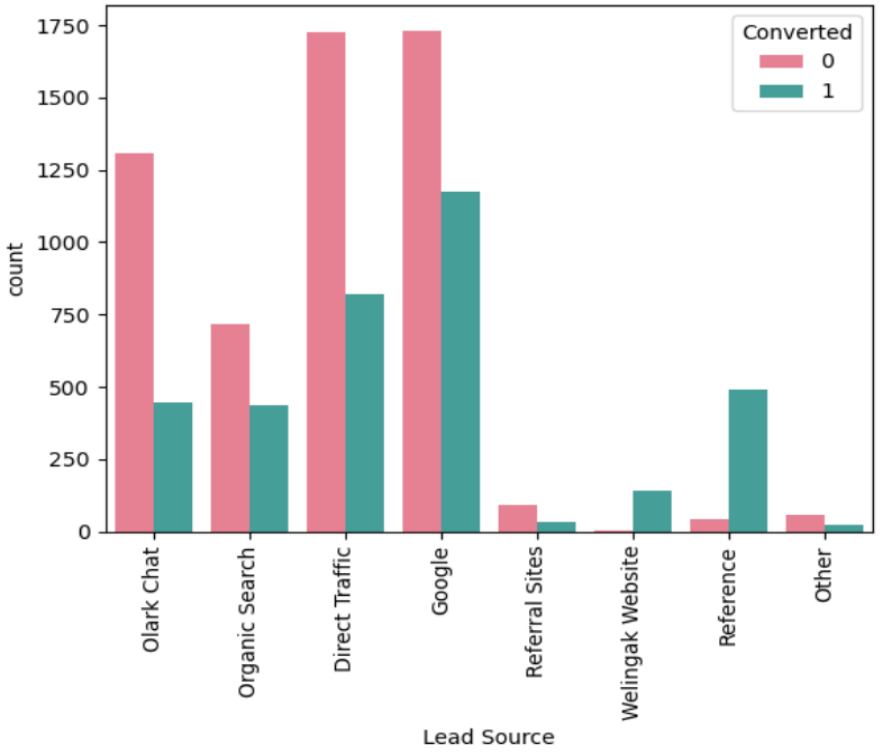
METHODOLOGY

Below are few of the considerations / assumptions we have made during this analysis:

- Many the features had a level called ‘Select’, we have imputed that with null values
- >35% of missing values have been dropped, we have also checked these features before dropping, we did not find them significant to our model
- We have created a new category (such as ‘Others’) in few features to combine lower valued data points or data points that had similar behaviour to avoid creating unnecessary dummy variables. We have also checked for duplicate data levels as well.
- Categorical variables with very less variance of data distribution (highly skewed data points) are either imputed or dropped based on their corresponding conversion rate.
- Numerical variables have gone through outlier treatment.
- Train and Test data sets are split with a ratio of 70-30, ‘Converted’ feature has become our ‘y’ variable while the rest are under X.
- We have used ‘Standard Scaler()’ to rescale the numeric features which aids in the total accuracy of model building.
- With the help of heatmap we have checked for multicollinearity between the variables to remove highly correlated variables to improve further efficiency of model.
- Using RFE with an estimator of ‘Logistic Regression()’ along with max of 20 steps we have further refined features.
- P-value has been checked for all features using the model summary, features with higher than 0.05 p-value have been dropped from the model, VIFs are also checked for models and are kept below 2.00
- Arbitrary cut off point of 0.5 was chosen initially which we then changed to 0.3 after plotting the sensitivity, specificity and accuracy plot along with ROC curve.
- Model is performing well with both test and train data sets, recall score or Sensitivity is above ~80% in both scenarios

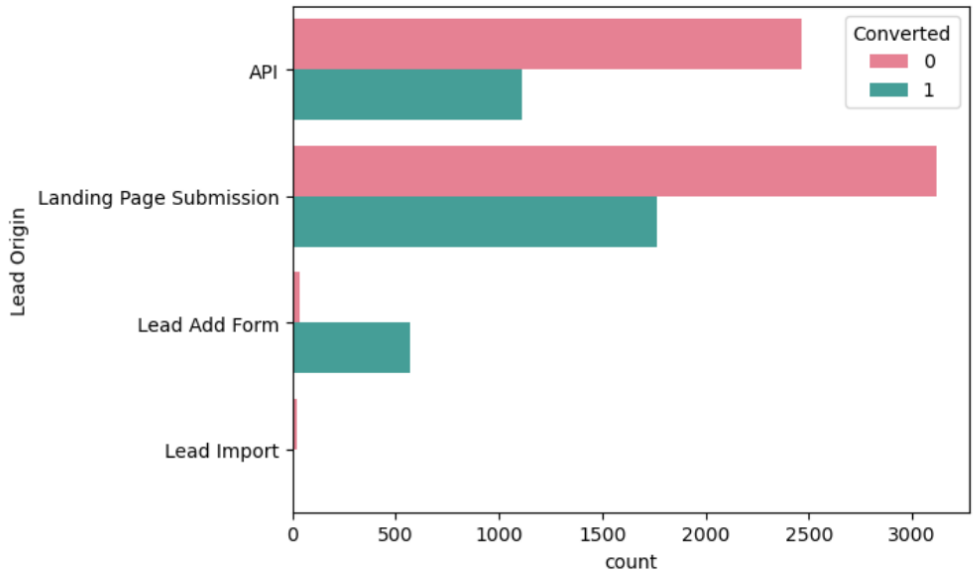
EDA

LEAD SOURCE



- Higher volume of leads are generated via 'Google' and 'Direct Traffic’.
- 'Reference' category has higher conversion rate even though the volume is lesser compared to other categories.
- While the volume is considerably higher in 'Olark chat', conversion rate does not look promising.

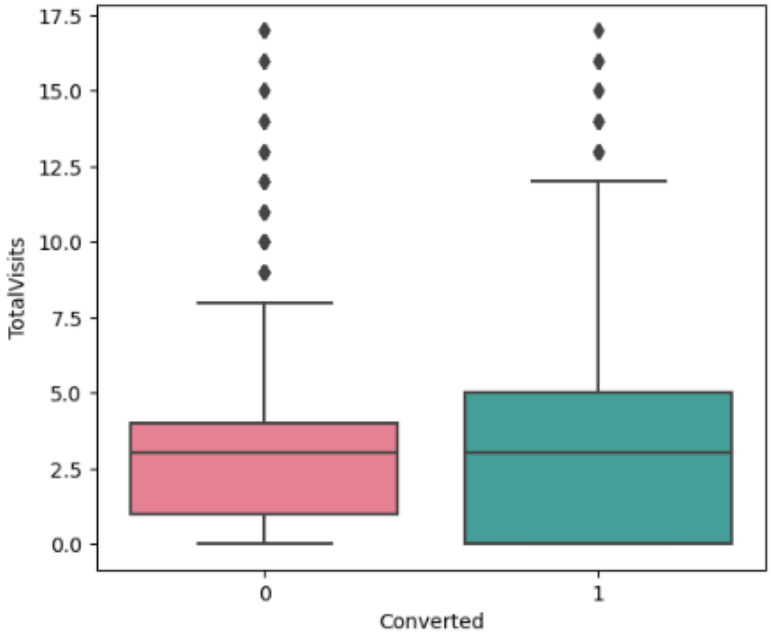
LEAD ORIGIN

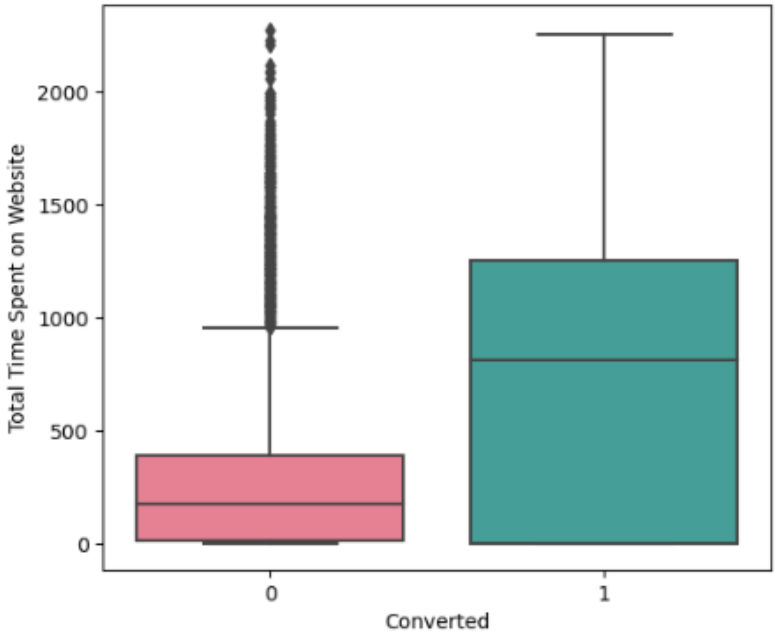


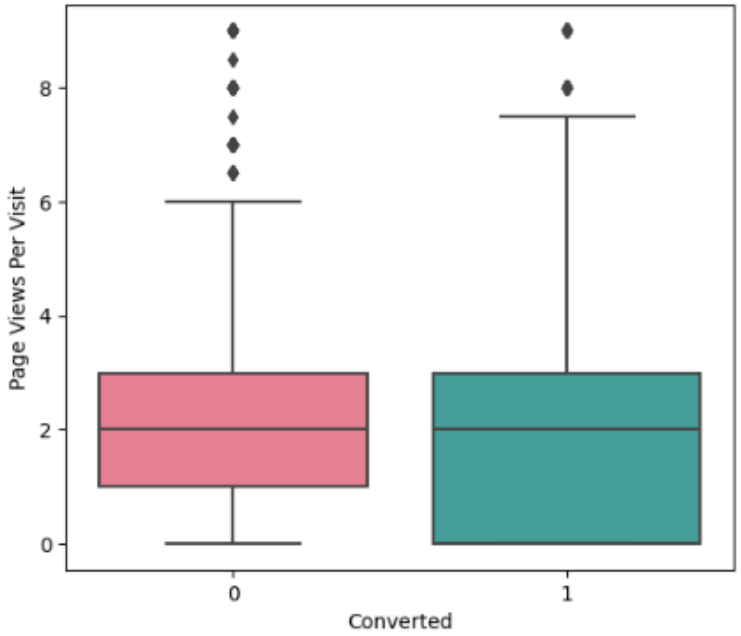
- 'Landing Page Submission' tops the charts with higher volume and higher conversion rate compared to other categories. We can also observe that 'Lead Add Form' is having higher conversion rate with a very less volume
- It is noticeable that we have a lot of room for improvement in 'Landing Page Submission' and 'API' with regards to lead conversion.

EDA – Numerical Variables

CONVERSION RATE AND DATA DISTRIBUTION OF ALL THE NUMERIC VARIABLES







- For 'Total Time Spent on Website' we can clearly witness that the conversion rate is higher for those who tend to spend more time on the website.
- However, for the other two variables the median looks to be very close in both cases of conversions.

MODEL SUMMARY

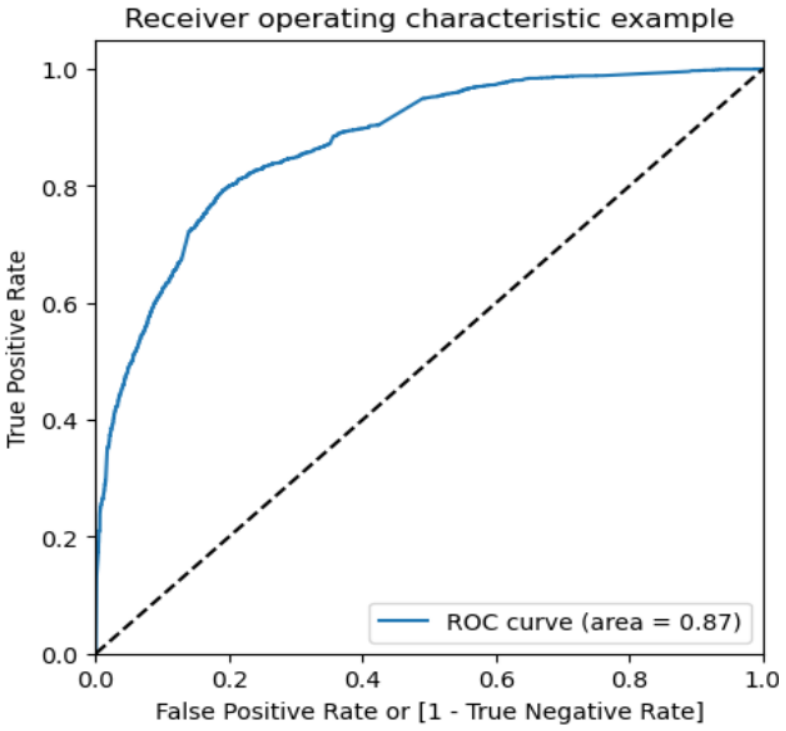
Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6267
Model:	GLM	Df Residuals:	6256
Model Family:	Binomial	Df Model:	10
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2679.2
Date:	Mon, 19 Jun 2023	Deviance:	5358.4
Time:	19:10:53	Pearson chi2:	6.47e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.3773
Covariance Type:	nonrobust		

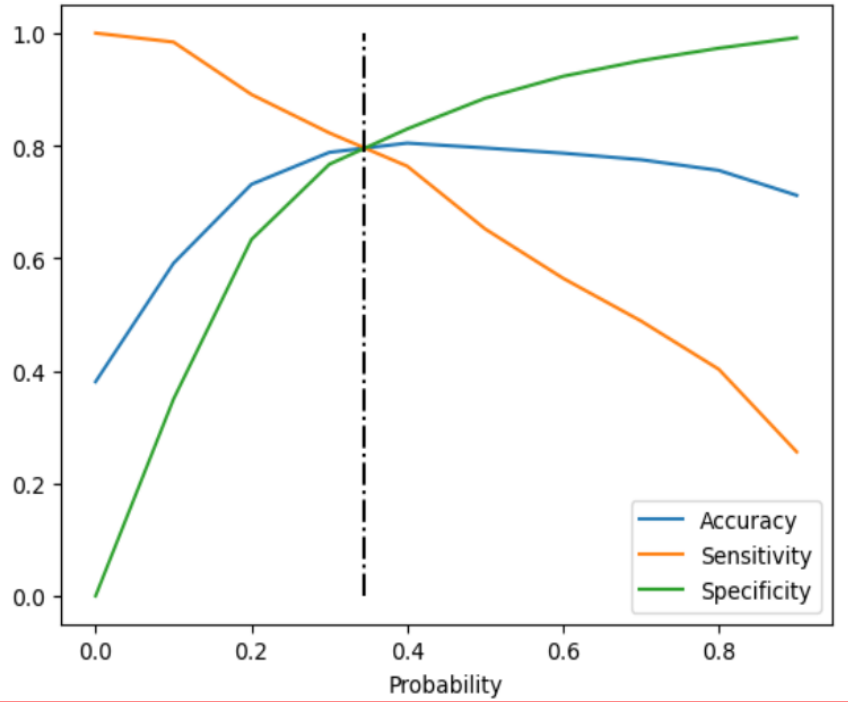
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.9858	0.082	-24.226	0.000	-2.146	-1.825
Do Not Email	-1.2555	0.165	-7.600	0.000	-1.579	-0.932
Total Time Spent on Website	0.9217	0.035	26.224	0.000	0.853	0.991
Lead Source_Direct Traffic	-0.6693	0.081	-8.262	0.000	-0.828	-0.511
Lead Source_Organic Search	-0.3769	0.105	-3.591	0.000	-0.583	-0.171
Lead Source_Reference	3.4441	0.258	13.335	0.000	2.938	3.950
Lead Source_Welingak Website	5.5428	1.010	5.485	0.000	3.562	7.523
Last Activity_SMS Sent	1.4356	0.073	19.753	0.000	1.293	1.578
What is your current occupation_Student	1.2099	0.237	5.100	0.000	0.745	1.675
What is your current occupation_Unemployed	1.2403	0.085	14.675	0.000	1.075	1.406
What is your current occupation_Working Professional	3.6891	0.203	18.151	0.000	3.291	4.087

MODEL EVALUATION

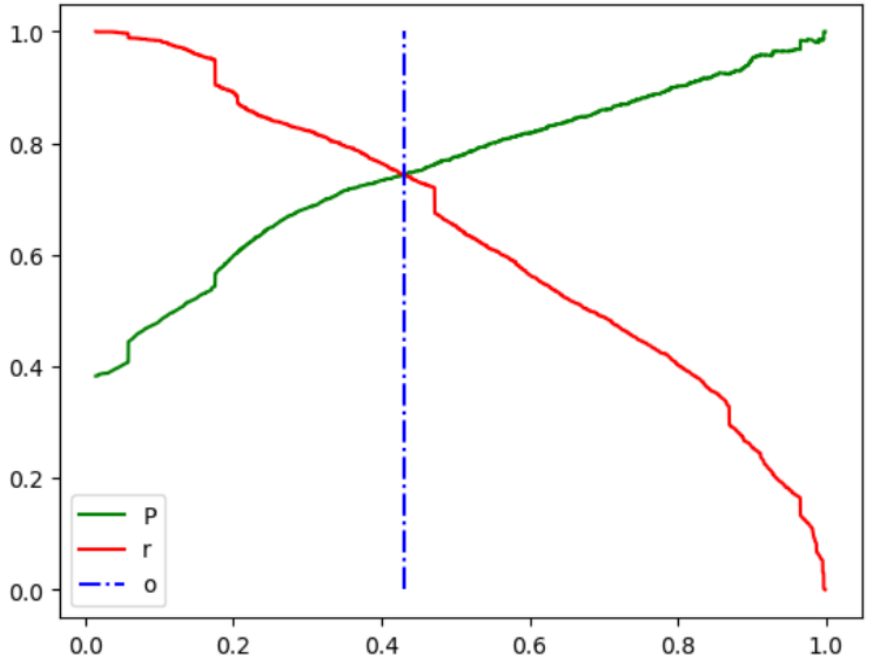
ROC CURVE



OPTIMAL CUT-OFF POINT



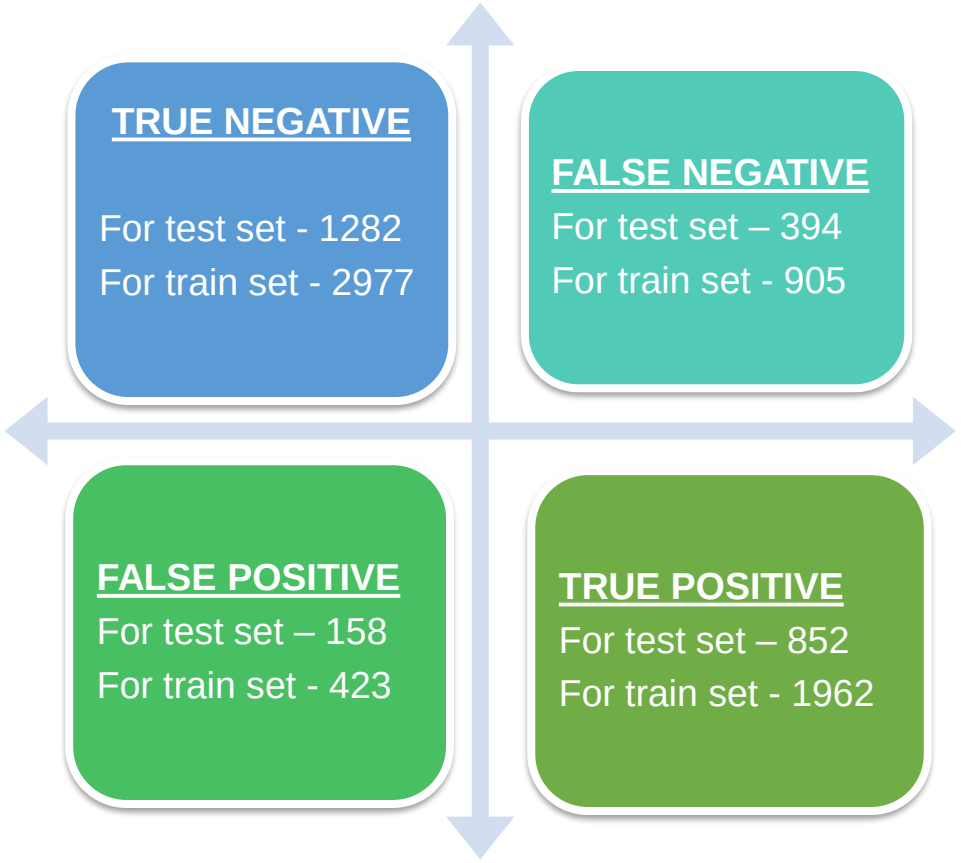
PRECISION AND RECALL



- ROC curve shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- **In future if the model needs modification then we can surely alter different ratios of sensitivity and specificity to obtain the desired results from the model.**
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- From the optimal cut off point curve above, 0.34 is the optimum point to take as a cutoff probability.
- In the model we have continued with 0.3 as we wanted higher sensitivity for the model.
- If we consider 'Precision and Recall' as our main evaluation metrics, then above curve indicates that the optimal cut-off point should be 0.43

MODEL RESULTS

CONFUSION MATRIX



EVALUATION METRICS RESULTS

TRAIN DATA SET:

- *Accuracy – 78.80%*
- *Sensitivity – 82.26%*
- *Specificity – 76.68%*
- *False Positive rate – 23.31%*
- *Positive Predictive Value – 68.43%*
- *Negative Predictive Value – 87.56%*
- *Precision – 77.59%*
- *Recall – 65.29%*

TEST DATA SET:

- *Accuracy – 79.94%*
- *Sensitivity – 84.36%*
- *Specificity – 76.39%*
- *False Positive rate – 23.51%*
- *Positive Predictive Value – 68.38%*
- *Negative Predictive Value – 89.02%*
- *Precision – 68.38%*
- *Recall – 84.36%*

CONCLUSION / SUGGESTIONS

- We can see that the final prediction of conversions are of 85% conversion, as per the X Educations CEO's requirement.

- Top features that are impacting the conversion rate positively are :
  - Lead sources originating from 'Welingak Website'
  - Customers that are working professionals tend to join the courses compared to other occupations
  - Lead sources originating via 'Reference'

- If the customer has opted 'Do not email' option they are most likely uninterested in the program and we can avoid wasting time on such leads.

- There are few areas where we see room for improvement in terms of conversion rate:
  - Under the feature 'Lead Origin' we have categories such as 'Landing Page Submission' and 'API', here even though we have a decent amount of leads volume they are not getting converted efficiently, team can dive deep into this.

  - Lead source is an important feature during our initial EDA it is found that both 'Google' and 'Direct Traffic' is generating very similar amount of lead volume however the 'Direct Traffic' leads' conversion is not up to the mark, perhaps sales team can probably find an alternative approach to these leads.

  - Sales team also has to relook at leads that are sourcing through 'Olark Chat' to understand if the process is user friendly or not because conversion rate is low compared to its volume.

SUBJECTIVE QUESTIONS WITH ANSWERS-1

1. Which are the top three variables in your model which contribute most towards the probability of a lead getting converted?

   Solution: Lead Source, Occupation of the customer and Last activity are the variables which are impacting the conversion rate positively.

2. What are the top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion?

   Solution: As per our model we have higher co-efficient for the following three features,

   a. **Lead Source_Welingak Website – 5.54**
   b. **What is your current occupation_Working Professional – 3.69**
   c. **Lead Source_Reference – 3.44**

SUBJECTIVE QUESTIONS WITH ANSWERS-2

3.X Education has a period of 2 months every year during which they hire some interns. The sales team, in particular, has around 10 interns allotted to them. So, during this phase, they wish to make the lead conversion more aggressive. So, they want almost all of the potential leads (i.e., the customers who have been predicted as 1 by the model) to be converted and hence, want to make phone calls to as much of such people as possible. Suggest a good strategy they should employ at this stage.

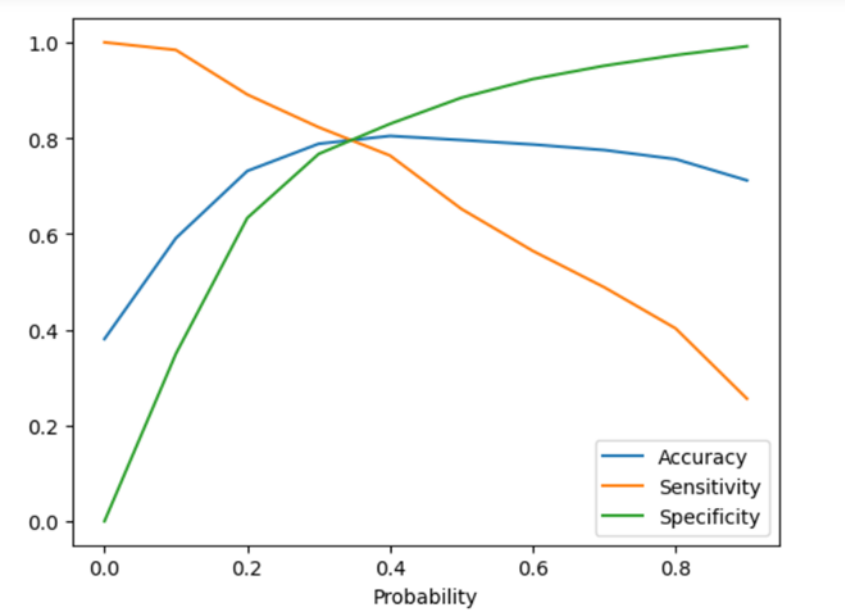
Solution: Every sale cycle will have its impact from seasonality. The first suggestion to management would be to,

- **To plan the hiring of these interns during the peak season (High volume of leads)**
- **Our model predicts 'Hot leads' with a recall score of 84.35% with an accuracy of ~80%, hence these leads have to be prioritized**
- **Recall score/Sensitivity is the percentage by which our model is predicting converted leads correctly over the total number of actual conversions.**
- **As shown in the graph below sensitivity can be regulated by using different optimal cut off point, if we need to target high conversion rate potential leads then we can find the cut off point which yields higher % of sensitivity.**

Next slide>>

SUBJECTIVE QUESTIONS WITH ANSWERS-3

- It is advisable that the sales team managers educate or provide the guidance to interns on important variables that have positive impact on the conversion rate
- Interns must also be aware of the red flags (Features that suggest negative impact on the conversion rate) as well so they can move on to another customer and avoid wasting time.



SUBJECTIVE QUESTIONS WITH ANSWERS-4

4. Similarly, at times, the company reaches its target for a quarter before the deadline. During this time, the company wants the sales team to focus on some new work as well. So, during this time, the company's aim is to not make phone calls unless it's extremely necessary, i.e., they want to minimize the rate of useless phone calls. Suggest a strategy they should employ at this stage.

Solution:

a. **In this scenario team can take help of another evaluation metric of the model which is 'Specificity'. This metric is contrary to Specificity and provides the negative conversion rate over total actual negatives.**
b. **Hence it is important to choose an optimal cut off point that yields in high specificity, this way sales team can avoid as many non-leads as possible.**
c. **There are chances that the model might also ignore the high potential leads in this case but that should not impact the business as the team has already reached its target for the quarter.**