LEAD SCORING CASE STUDY USING LOGISTIC REGRESSION

By,

Tanuja Modupalli
Mrinmayi Kunkalienca
Mrunal Paunikar

CONTENTS

- Problem Statement
- Business objective
- Problem Approach
- EDA
- Test- Train split
- Correlations
- Model Evaluation
- Observations
- Summary
- Conclusion
- Recommendations

PROBLEM STATEMENT

- An education company named X Education sells online course to industry professionals. On any given day, many professionals who are interested in the course land on their website & browse for courses. They have the process of collecting data by form filling on their website, after which company decides to follow that individual as a lead.
- Once these leads are acquired, employees from sales team start making calls, writing mails etc., Through this process some get converted while some don’t
- Typical lead conversion at X education is 30%, which means if there are 100 people only 30 get converted and enroll to their program. To make this process more efficient company wishes to identify most potential leads known as “HOT LEADS”.
- If they successfully identify this set of leads, lead conversion rate should go up as sales team will now exclusively focus on these HOT LEADS rather than calling and communicating with every single member on the list.

BUSINESS OBJECTIVE

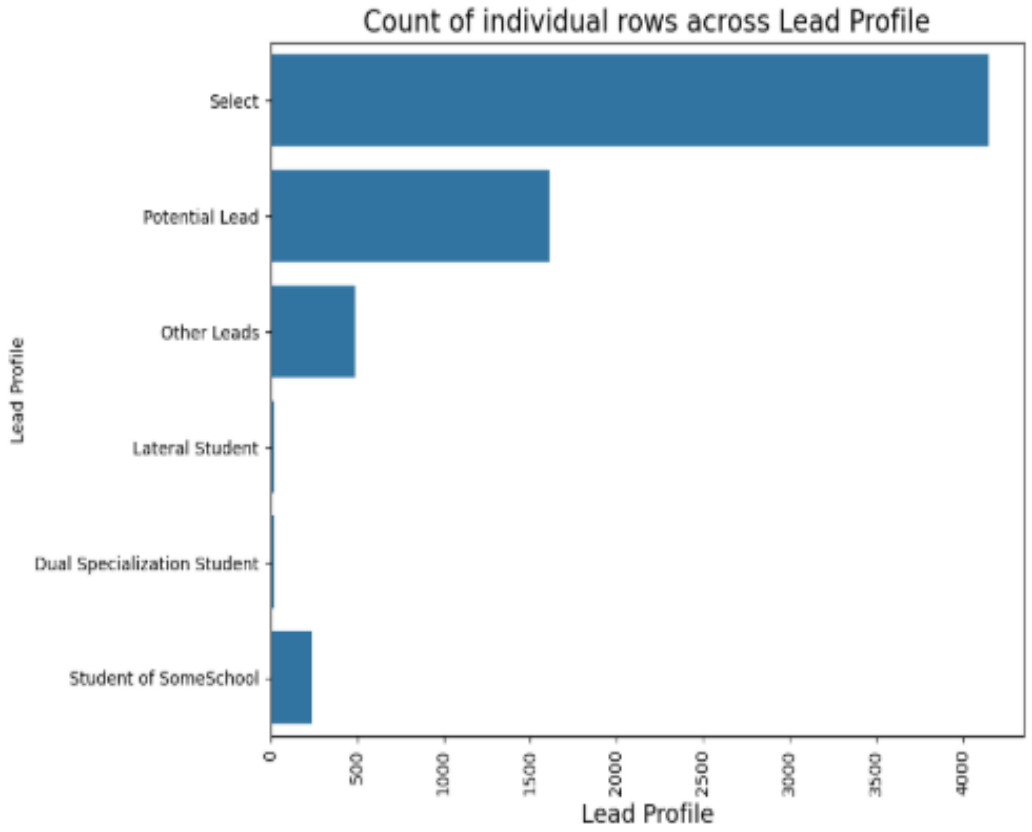
- Lead X wants us to build an model to give every lead a lead score between 0-100.
  so that they can identify the HOT LEADS and increase their conversion rate as well.
- The CEO wants to achieve a lead conversion rate around 80%
- They want the model to be able to handle future constraints as well as peak time actions required, how to utilize full man power and after achieving target what should be the approach

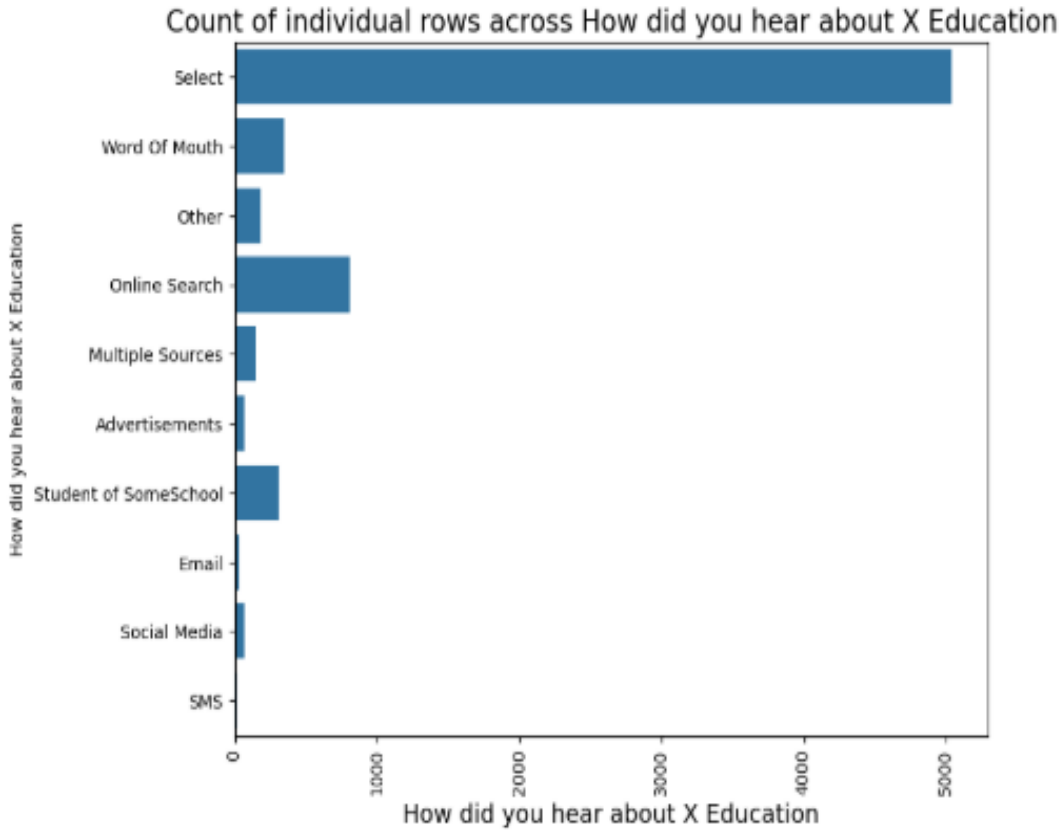
PROBLEM APPROACH

- Importing the data and inspecting the data frame
- Data preparation
- EDA
- Dummy variable creation
- Test-Train Split
- Feature scaling
- Correlations
- Model Building (RFE squared, VIF and p-values)
- Model Evaluation
- Making predictions on data set.

EDA DATA CLEANING

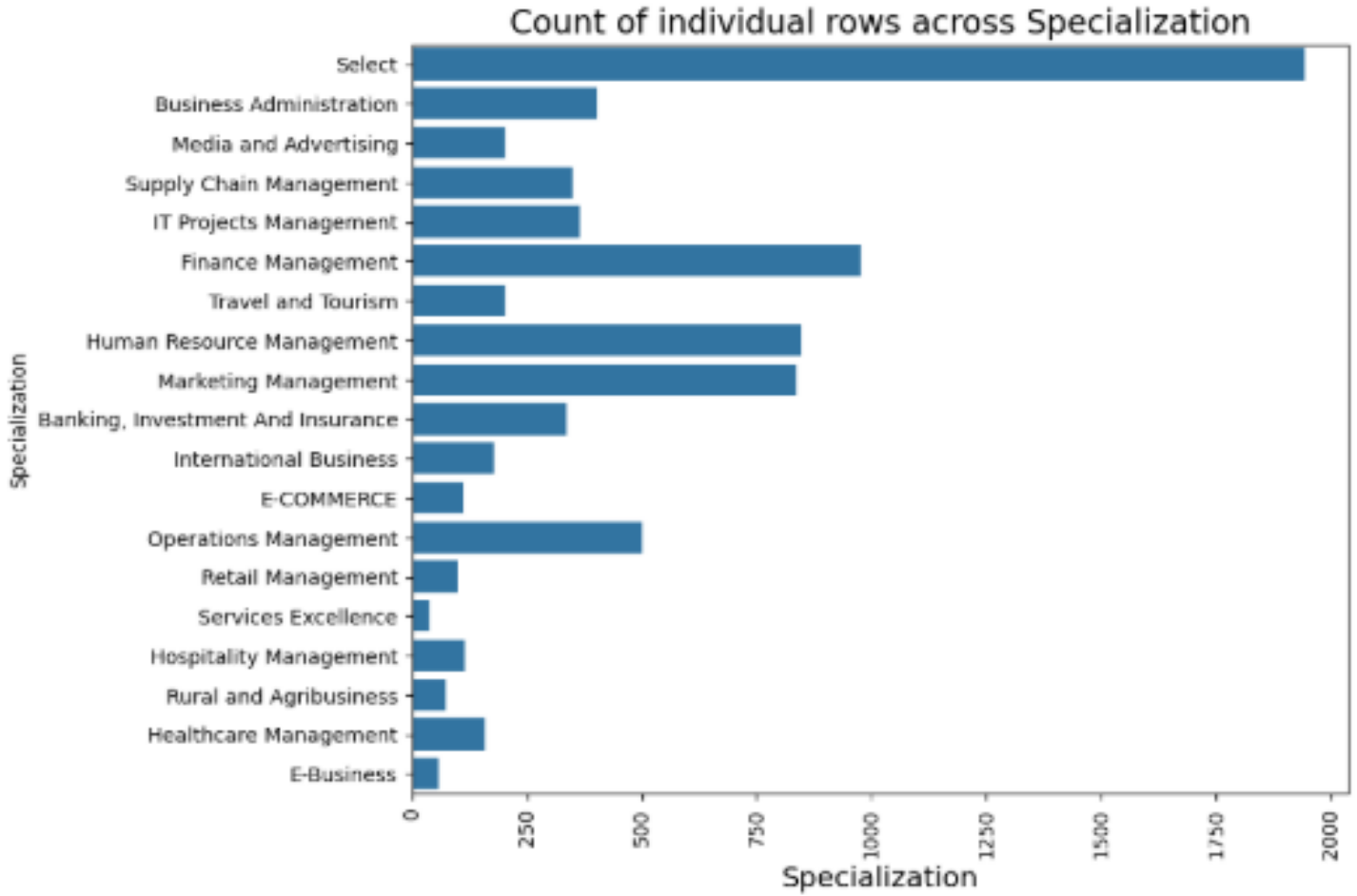
There are a few columns which have the attribute SELECT on which we have plotted graphs for LEAD PROFILE, How did you hear about X education and specialization





EDA DATA CLEANING

From below graph on specialization, we can see that LEADs from HR, Finance and Marketing management specializations have high probability to convert



TEST – TRAIN SPLIT

Here we have done a test train split in the ratio of 70-30

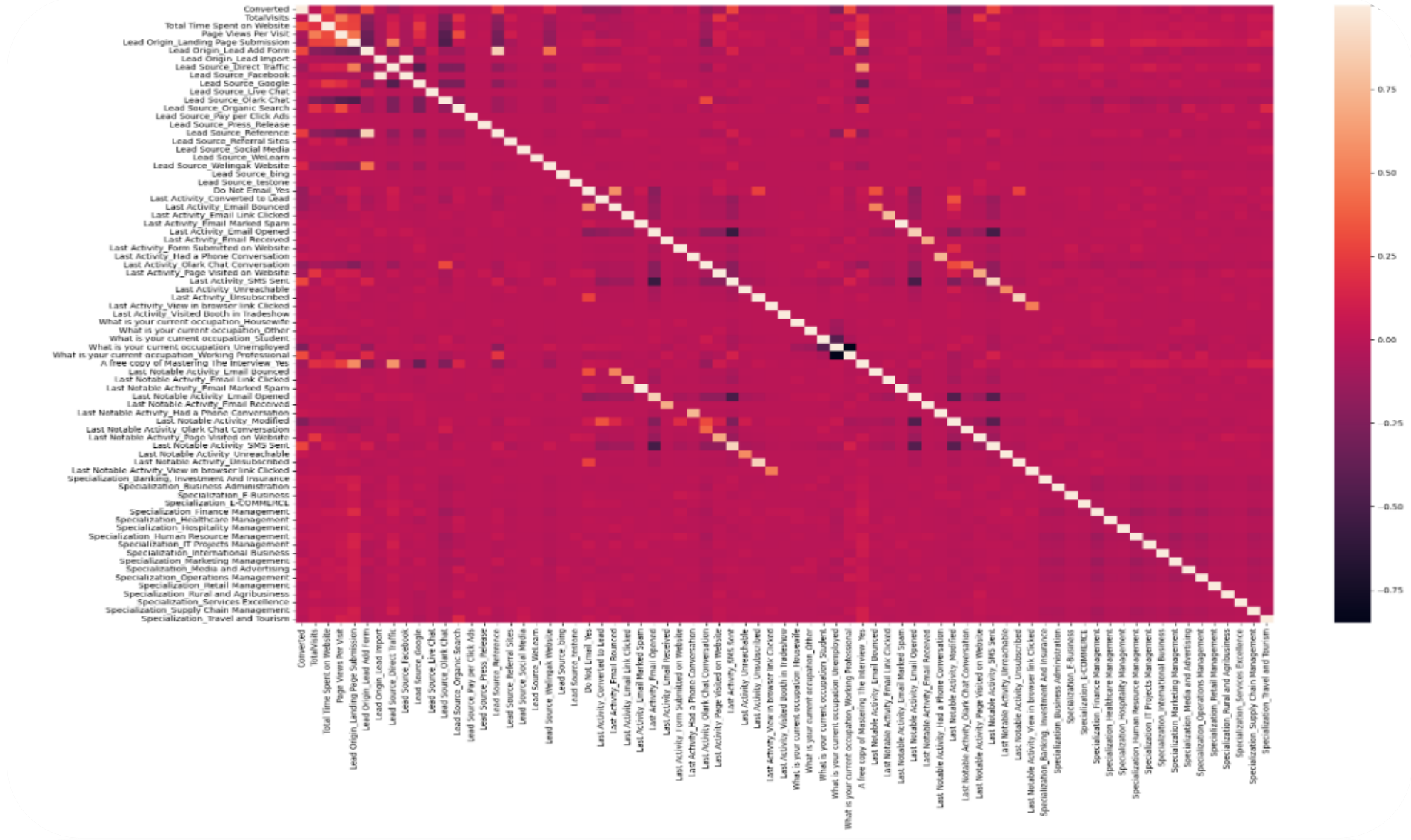
70% of data is trained and then testing is done on the remaining 30% data

```
# Split data:


X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.7, test_size=0.3, random_state=100)
```

```
# Lets check the shape of X_train and Y_train


print("X_train Size", X_train.shape)
print("y_train Size", y_train.shape)
```

```
X_train Size (4461, 74)
y_train Size (4461,)
```

CORRELATIONS



MODEL EVALUATION – FINAL MODEL

Generalized Linear Model Regression Results

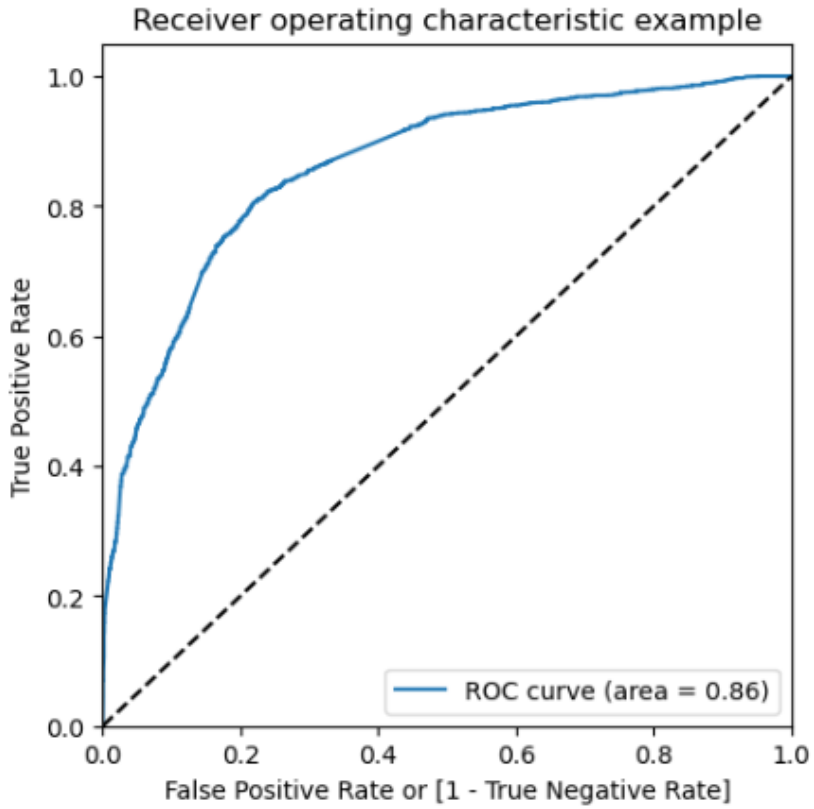
Dep. Variable:	Converted	**No. Observations:**	4461
Model:	GLM	**Df Residuals:**	4449
Model Family:	Binomial	**Df Model:**	11
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2079.1
Date:	Mon, 17 Feb 2025	**Deviance:**	4158.1
Time:	21:52:19	**Pearson chi2:**	4.80e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.3642
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2040	0.196	1.043	0.297	-0.179	0.587
TotalVisits	11.1489	2.665	4.184	0.000	5.926	16.371
Total Time Spent on Website	4.4223	0.185	23.899	0.000	4.060	4.785
Lead Origin_Lead Add Form	4.2051	0.258	16.275	0.000	3.699	4.712
Lead Source_Olark Chat	1.4526	0.122	11.934	0.000	1.214	1.691
Lead Source_Welingak Website	2.1526	1.037	2.076	0.038	0.121	4.185
Do Not Email_Yes	-1.5037	0.193	-7.774	0.000	-1.883	-1.125
Last Activity_Had a Phone Conversation	2.7552	0.802	3.438	0.001	1.184	4.326
Last Activity_SMS Sent	1.1856	0.082	14.421	0.000	1.024	1.347
What is your current occupation_Student	-2.3578	0.281	-8.392	0.000	-2.908	-1.807
What is your current occupation_Unemployed	-2.5445	0.186	-13.699	0.000	-2.908	-2.180
Last Notable Activity_Unreachable	2.7846	0.807	3.449	0.001	1.202	4.367

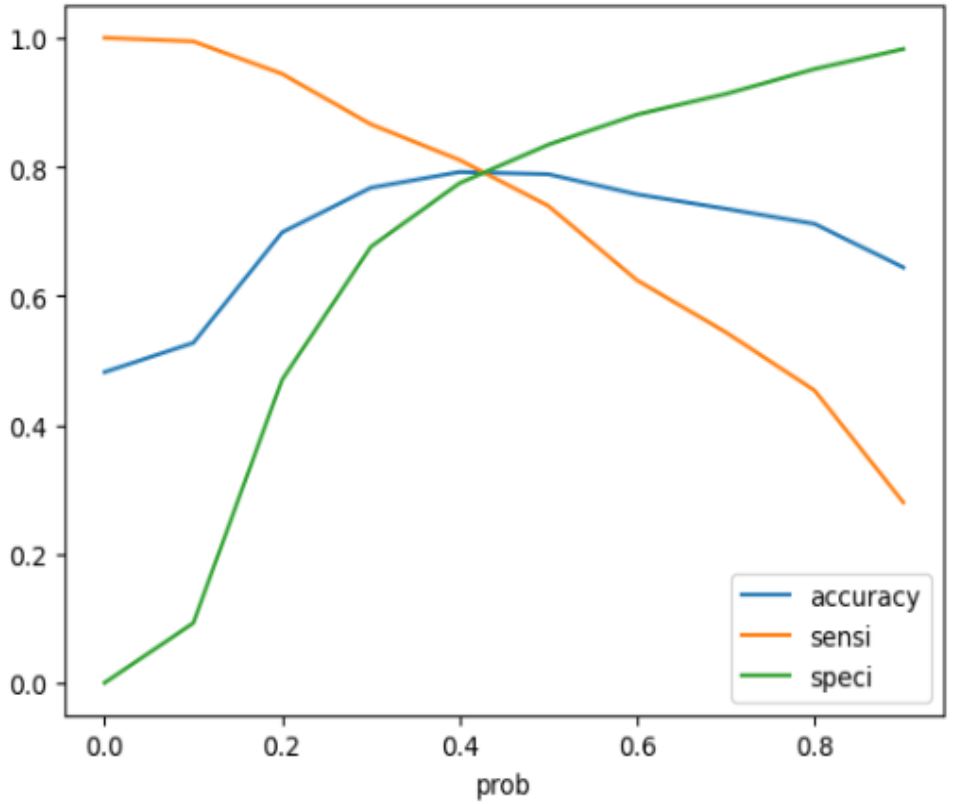
	Features	VIF
9	What is your current occupation_Unemployed	2.82
1	Total Time Spent on Website	2.00
0	TotalVisits	1.54
7	Last Activity_SMS Sent	1.51
2	Lead Origin_Lead Add Form	1.45
3	Lead Source_Olark Chat	1.33
4	Lead Source_Welingak Website	1.30
5	Do Not Email_Yes	1.08
8	What is your current occupation_Student	1.06
6	Last Activity_Had a Phone Conversation	1.01
10	Last Notable Activity_Unreachable	1.01

MODEL EVALUATION – ROC CURVE

We have ROC curve area as 0.86 and optimal cutoff as 0.41 as shown in below graphs



The area under the curve of the ROC is 0.86 which is quite good. So we seem to have a good model.



around 0.41 we have optimal values for all 3 params

OBSERVATIONS

1.Train Data:
- Accuracy: 79.197%
- Recall / Sensitivity : 79.944%
- Specificity: 78.053%

2. Test Data:
- Accuracy: 78.34%
- Recall / Sensitivity : 78.8209 ~ 79%
- Specificity: 77.91%

3. Optimal Cutoff considered: 0.41

4. Area under ROC: 0.86

SUMMARY

- Lead scoring case study has been done using Logistic Regression model to meet constraints as per business model.
- There are a lot of leads in the initial stage but only a few of them got converted into paying customers. Majority of the leads were from INDIA, MUMBAI.
- There were few columns with the attribute SELECT which basically meant that the Student had not selected the option for that particular column which is why it shows ‘Select’. To get some useful data we have to make some compulsory decision either to retain them or drop them based on the columns in which this was present.
- The highest number of total visits and total time spent on Website may increase the chances of lead to be converted.
- The leads have joined course for better career prospects as we saw that majority of them were from HR, Marketing and Finance specializations which had a higher chance to convert.
- Talking or communicating with notable Activity, making improvement in customer engagement through emails and calls will help convert leads. As the leads who open EMAIL have high chance of converting same as sending SMS will also benefit.
- Most leads current occupation is Unemployed, which gave more focus on Unemployed leads.

CONCLUSION

Based on the logistic regression model, the top three variables that contribute most towards the probability of a lead getting converted are:

- **Total Time Spent on Website:** This variable has a significant positive impact on the likelihood of conversion. Leads that spend more time on the website are more likely to convert, as they show a higher level of interest in the courses.

- **Last Activity_Had a Phone Conversation:** Having a phone conversation significantly increases the chance of conversion. This suggests that direct interaction with leads can be a powerful tool in moving them further down the funnel.

- **Lead Origin_Lead Add Form:** Leads generated through the lead add form on the website have a higher conversion rate. These leads likely have a clearer intent to engage with the company, which makes them more likely to convert.

RECOMMENDATIONS

The top three categorical/dummy variables to focus on in order to increase the probability of lead conversion are:

- **Lead Source_Olark Chat:** Leads coming from the Olark chat source have a higher chance of conversion. This suggests that real-time communication with leads through chat is an effective way to nurture them towards conversion.

- **Last Activity_Had a Phone Conversation:** Leads who have had a phone conversation with the sales team have a higher conversion probability. This emphasizes the importance of personal contact in converting leads into customers.

- **What is your current occupation_Student:** Interestingly, leads who are students have a lower probability of conversion. Focusing on reducing this demographic's involvement or adjusting the sales approach for them may increase overall conversion rates.

THANK YOU! ☺

- Tanuja Modupalli
  Mrinmayi Kunkaliencar
  Mrunal Paunikar