LEAD SCORING CASE STUDY

Vishnuvardhan Sanem & Tuhin Mondal

DS 40 – 18th July 2022

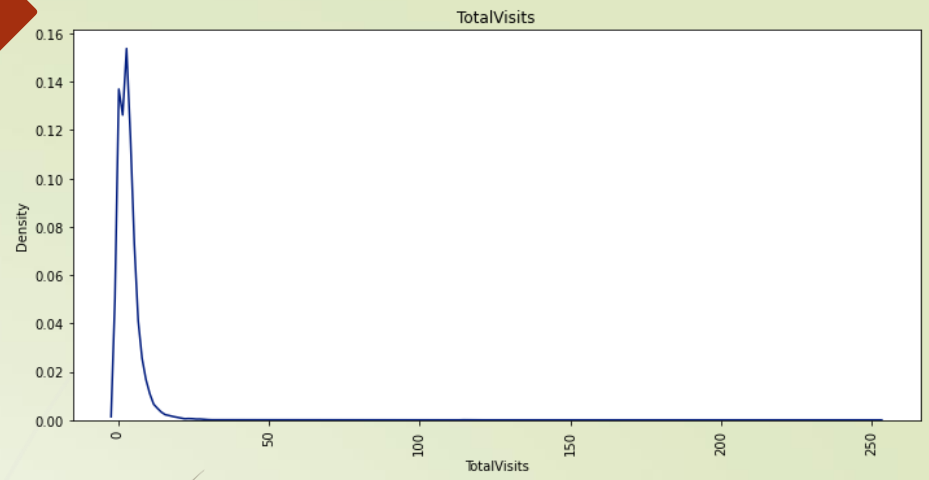
PROBLEM STATEMENT

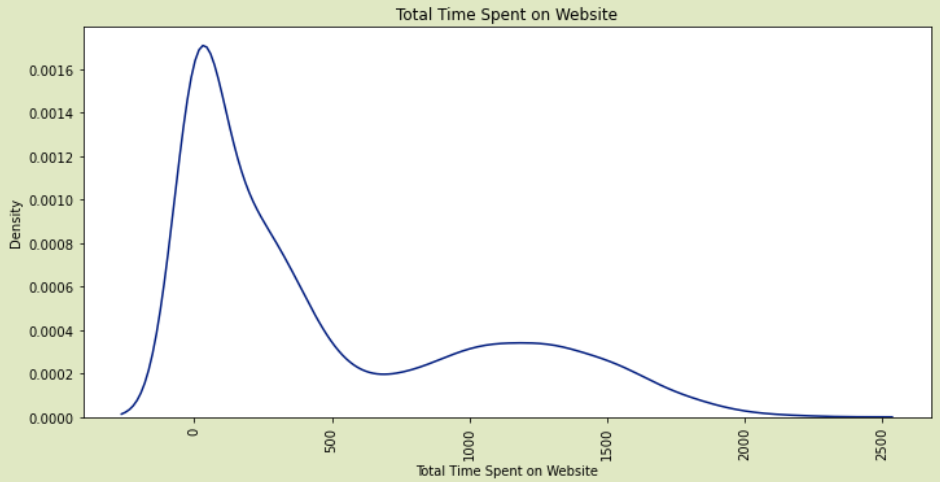
- Our company, X Education markets its online courses on several websites and search engines like Google.
- People are termed as a lead when they fill up a form providing their email address and phone numbers.
- The Sales team then contacts them via phone calls and mails and convince them to take up an online course through our platform.
- The typical conversion rate is only 30%.
- Now, our company wants to identify its potential customers so that the Lead conversion rate gets higher and effort and the cost that goes behind every lead gets optimized.
- We want to focus on our Hot Leads by engaging them in more conversations and offering discounts and explaining the benefits in detail.
- We need to generate a score which will help us identify these Hot Leads and target them aggressively instead of focusing on a large cohort.
- Our time and effort get valued in this way and remove the casual customers from our target. **Our Objective is to build a robust model generating Lead Scoring for every customer and providing the Sales team with a list of the Hot Leads.**

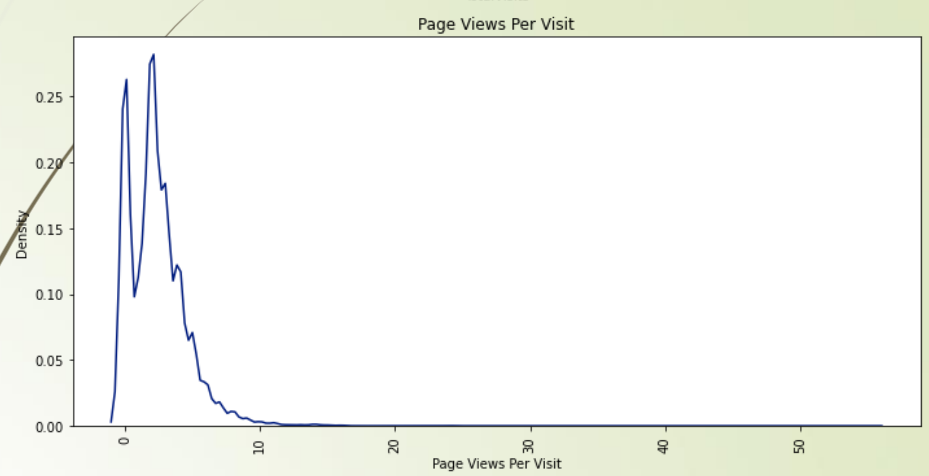
ANALYSIS APPROACH

- We perform some basic data cleaning steps like missing value imputation and grouping of fields with low counts in a particular variable.
- We drop some variables based on their skewness towards a category and then treat for outliers.
- We perform some EDA on the categorical and continuous variables and recommend some measures to try and increase the conversion rates based on those.
- We build a logistic regression model with 20 selected variables from RFE and remove some of them from the model based on their p-values and VIF.
- We get a model with 14 feature variables explaining our target variable, 'Converted'.
- As the next most important step of the analysis, we assign a Lead Score to each lead using the conversion probability.
- We check for the evaluation metrics like sensitivity, specificity, precision after we select an optimal cutoff point using ROC and then validate our results on the test set.
- The training and the test sets show results on similar lines and our model will serve as a key guide to X Education and their business goals.

EXPLORATORY DATA ANALYSIS



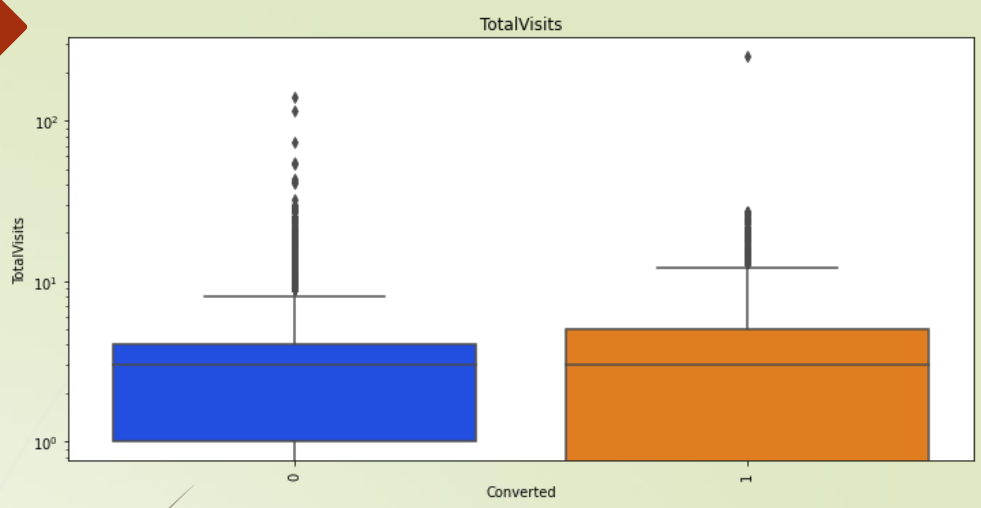


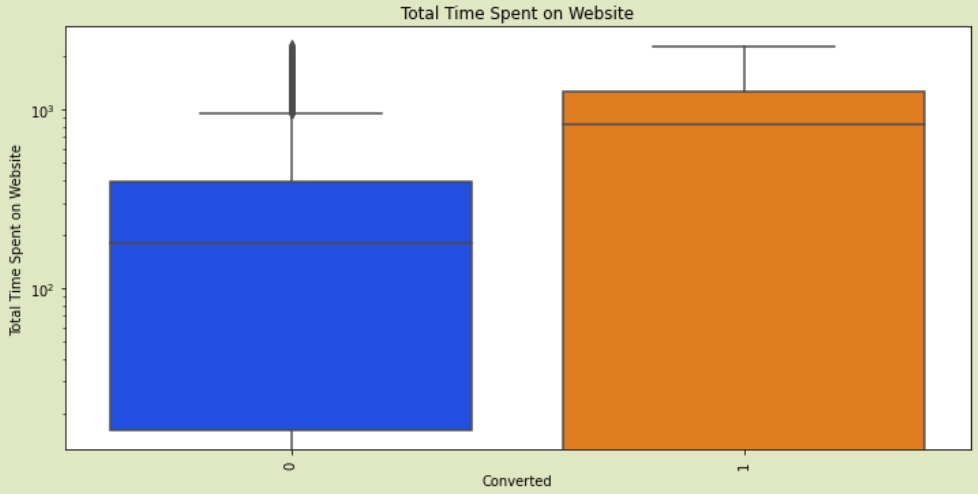


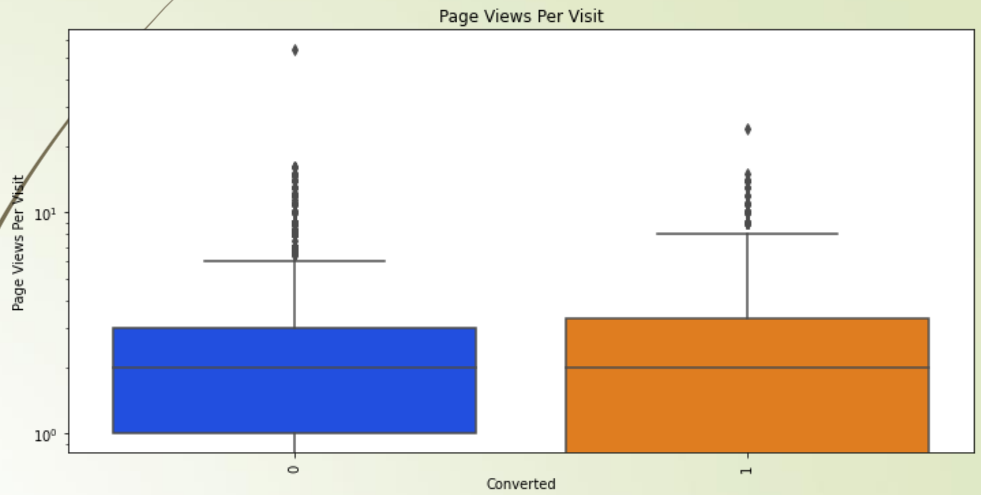
Uni-variate Analysis:

- The max probability for TotalVisits is found to be around 15-20. It increases initially but decreases further.
- The max probability for PageViewsPerVisit is found to be around to be 3-5
- The probability of time spent is found to be high for time between 0-300 seconds and decreases further.

EXPLORATORY DATA ANALYSIS



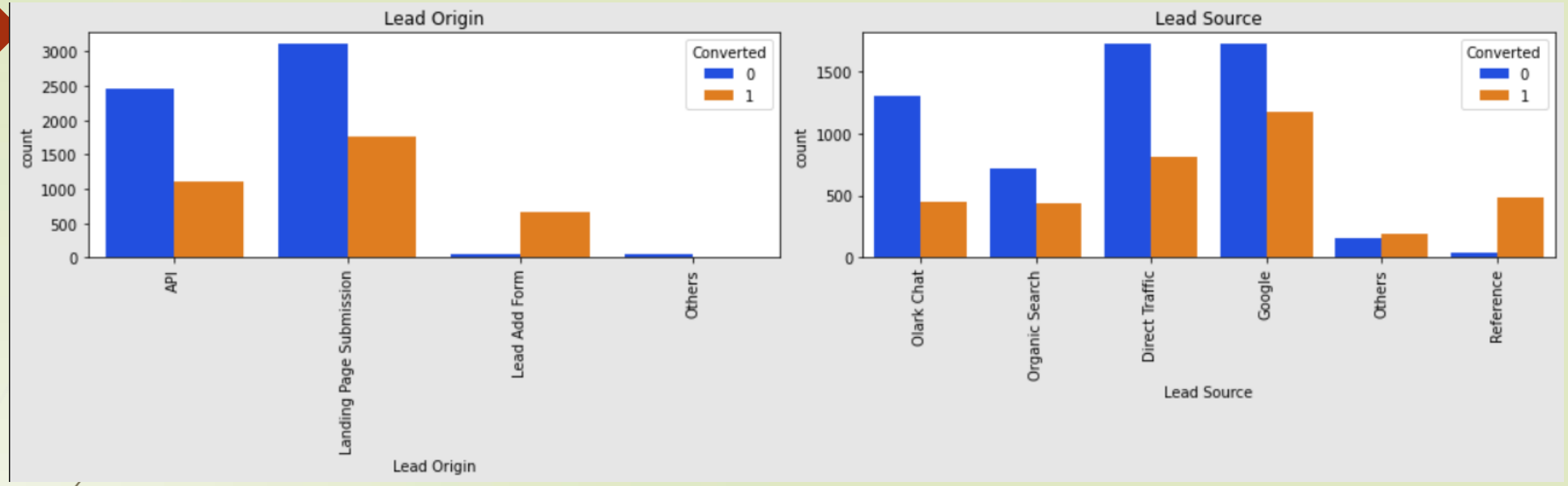


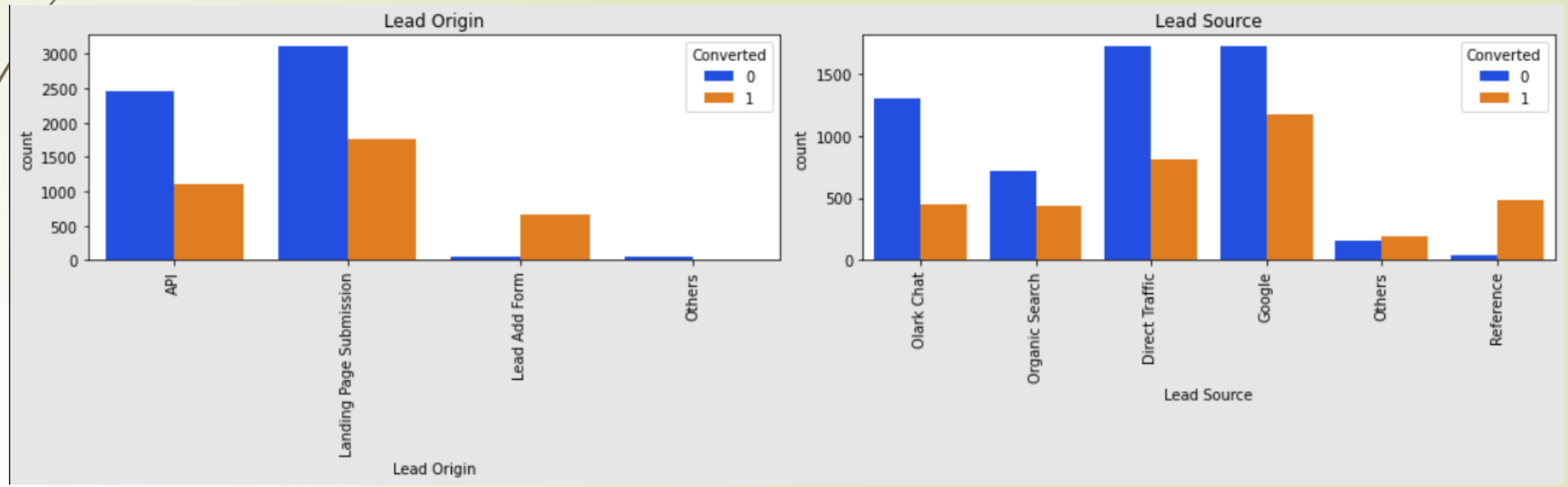


Bi-variate Analysis:

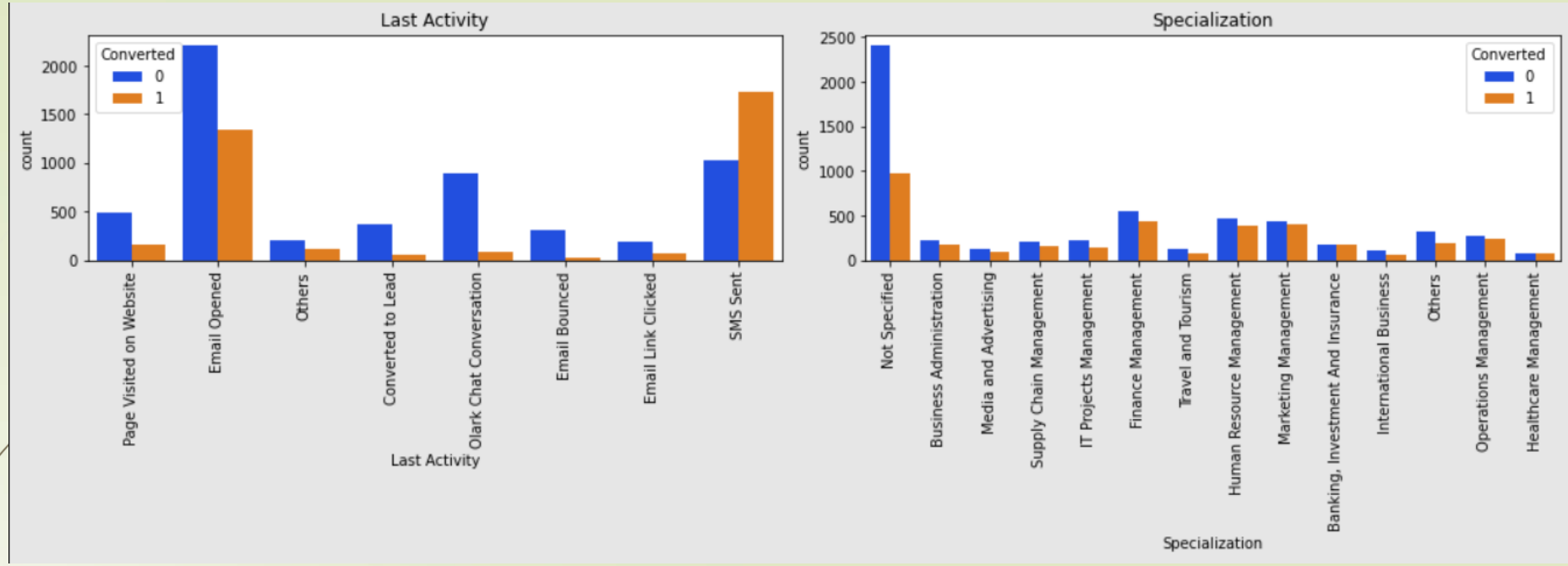
- The mean is found to be higher in case of Converted people rather than non-converted people.
- The average page views for both converted and non converted is found to be the same.
- The average total visits for both converted and non converted people is found to be the same.

EXPLORATORY DATA ANALYSIS





EXPLORATORY DATA ANALYSIS



Univariate analysis of Categorical data

•The % of converted people is found to be greater for Landing Page Submission. We can also see that if Lead source is Add Form, the ratio of lead conversion is very high(almost not converted is very less).

•Google is found to be the important source for Lead Conversion

•It is clearly visible from the graph that we need to target the Unemployed and Working Professional to get a higher conversion rate. The ratio of conversion rate is higher than not converted people for working professionals.

•We need to target people via Emails and SMS as it is found that the probability of response in case Converted leads is found to be higher.

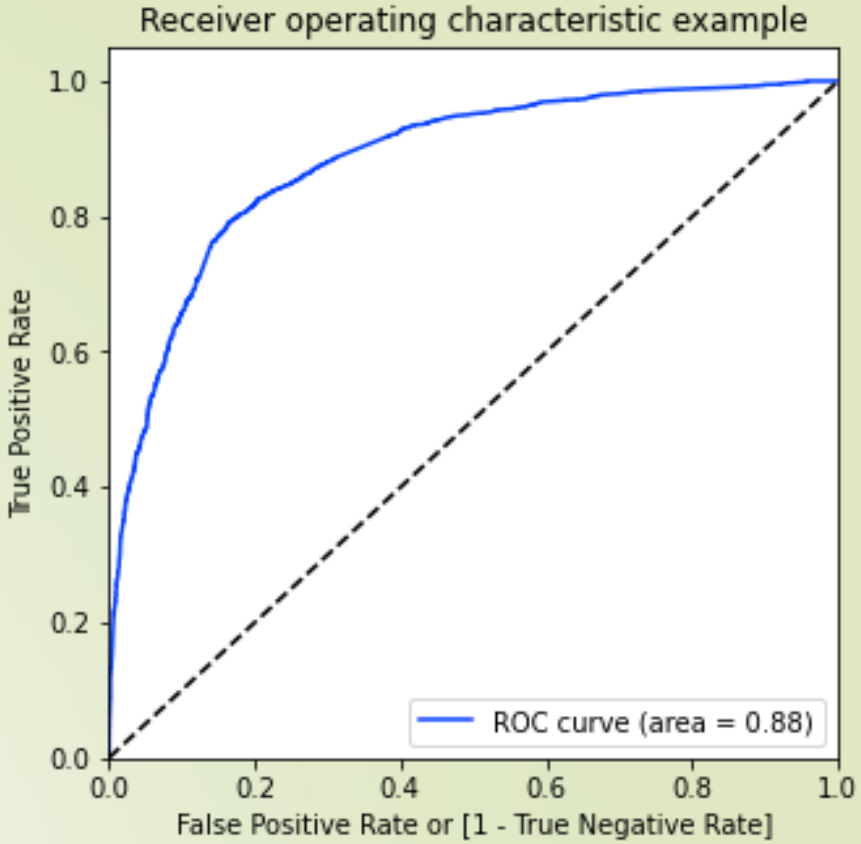
•We cannot infer much about conversion rate from specialization as people who do not select any specialization can also be converted to a lead but the ratio of non converted leads is higher than converted ones if they didn't choose specialization.

•People usually don't subscribe for a free copy of mastering the interview.

Features of the Final Model

	coef
const	-1.0565
TotalVisits	0.1944
Time Spent	1.0574
Free Copy	-0.3186
Lead Origin_Landing Page Submission	-1.0199
Lead Origin_Lead Add Form	4.4017
Lead Source_Olark Chat	1.2101
Lead Source_Reference	-1.1764
Last Activity_Email Bounced	-1.1921
Last Activity_Email Opened	0.8166
Last Activity_Olark Chat Conversation	-0.6859
Last Activity_Others	0.6463
Last Activity_SMS Sent	1.9097
Specialization_Not Specified	-1.1380
Current Occupation_Working Professional	2.6908

ROC CURVE



- The ROC Curve shows us trade off between sensitivity and specificity.
- The Area Under ROC Curve or AUROC is 0.88, which is pretty high and bears testimony to the high accuracy and predictive power of our model.

EVALUATION METRICS

METRICS	TRAIN DATA	TEST DATA
Accuracy	81.20%	80.09%
Sensitivity	80.45%	80.00%
Specificity	81.66%	80.38%
Precision	72.99%	72.61%

- We have evaluated our model's prediction power on several metrics and all of them have come up with very good scores. Moreover, these metrics show us a similar results on both the Train and Test Data, implying the robustness and high predictive power of our model.

FINAL RESULT

	Converted	Lead_Number	Conversion_Prob	final_predicted	Score
2656	1	634047	0.999681	1	99.97
3478	1	627106	0.999659	1	99.97
8074	1	588037	0.999566	1	99.96
3428	1	627462	0.999458	1	99.95
5921	1	604411	0.999389	1	99.94

- We have generated **Lead Scores** for every Prospect ID, the unique identification key for a lead. We can use this Lead Score to identify a Hot Lead. A cutoff point, say around 80, should help us in the X Education to contact a Lead with the highest chances of getting converted. We can ignore the people with lower Lead Scores and save our Sales team's time and X Education's money.

CONCLUSION:

•Conversion Rate for hot leads is increases from 73% to 96%. This means they have a 96% probability of getting converted to a lead.

•Focusing on Hot Leads will increase the chances of obtaining more value to the business as the number of people we contact are less but the conversion rate is high.

From our model, we can conclude the below points :

•customers who fills the form are most likely the potential leads.

•majorly focus on working professionals.

•majorly focus on leads whose last activity is SMS sent or Email opened.

•If the lead source is referral, he/she may not be the potential lead.

•If the lead didn't fill specialization, he/she may not know what to study and are not right people to target.

•Its always good to focus on customers, who have spent more time on our website.

•Its better to focus least on customers to whom the sent mail is bounced back.

RECOMMENDATIONS:

•Its good to collect data often and run the model and get updated with the potential leads. There is a belief that the best time to call your potential leads is within few hours after the lead shows interest in the courses.

•Along with phone calls, it's good to mail the leads also to keep them reminding as email is as powerful as cold calling.

•Reducing the number of call attempts to 2-4 and increasing the frequency of usage of other media like advertisements in Google, or via emails to keep in touch with the lead will save a lot of time.

•Focusing on Hot Leads will increase the chances of obtaining more value to the business as the number of people we contact are less but the conversion rate is high.