# Lead Scoring Case Study

Abdulhadi Kanjo
Shruthi S Nayak
Parmesh sharma

upGrad & IIITB | Data Science Program - Nov 2023

# Business Problem Statement

- X Education sells online courses to industry professionals.
- 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.

# Business Objective

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

# Solution Methodology

- Data Cleaning :-
  - Dropping columns that have only one unique values for all the leads.
  - Handling 'Select' variable that is present in many categorical variables.
  - Dropping all the columns with more than 40% missing values
  - Dropping columns with high data imbalance
  - Using imputation technique for columns having less % of missing values
  - Combining categories having low percentages into one single category

# Solution Methodology

- EDA :-
  - Univariate data analysis: value count, distribution of variable etc.
  - Bivariate data analysis: correlation coefficients and pattern between the variables etc.
- Data Preparation :-
  - Converting binary variable (Yes/No) to 1/0.
  - Creating dummies for categorical columns
  - Performing train-test split : The split was done at 70% and 30% for train and test data respectively.
  - Performing Scaling :We'll do the standard scaling on the variables ['TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website']
  - Finding correlations between different variables using heat map.

# Solution Methodology

- Model Building
- Classification Technique: logistic regression used for the model making and prediction.
- RFE to be used for model feature selection.
- After RFE , P- values and VIFs to be checked.
- Confusion Matrix
  - Validation of the model.
  - Sensitivity – Specificity
  - Precision – Recall
- Model presentation
  - Conclusions and recommendations

# Data Cleaning

- Total Number of Rows =37,
- Total Number of Columns =9240.
- Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.
- Dropping Prospect ID as these are just indicative of the ID number of the Contacted People.
- Inspecting all the columns with Select variable in the dataframe. ‘Select’ values replaced with null values in columns like ‘Specialization’ , 'How did you hear about X Education’ , 'Lead Profile’ , ‘City’.

# Data Cleaning

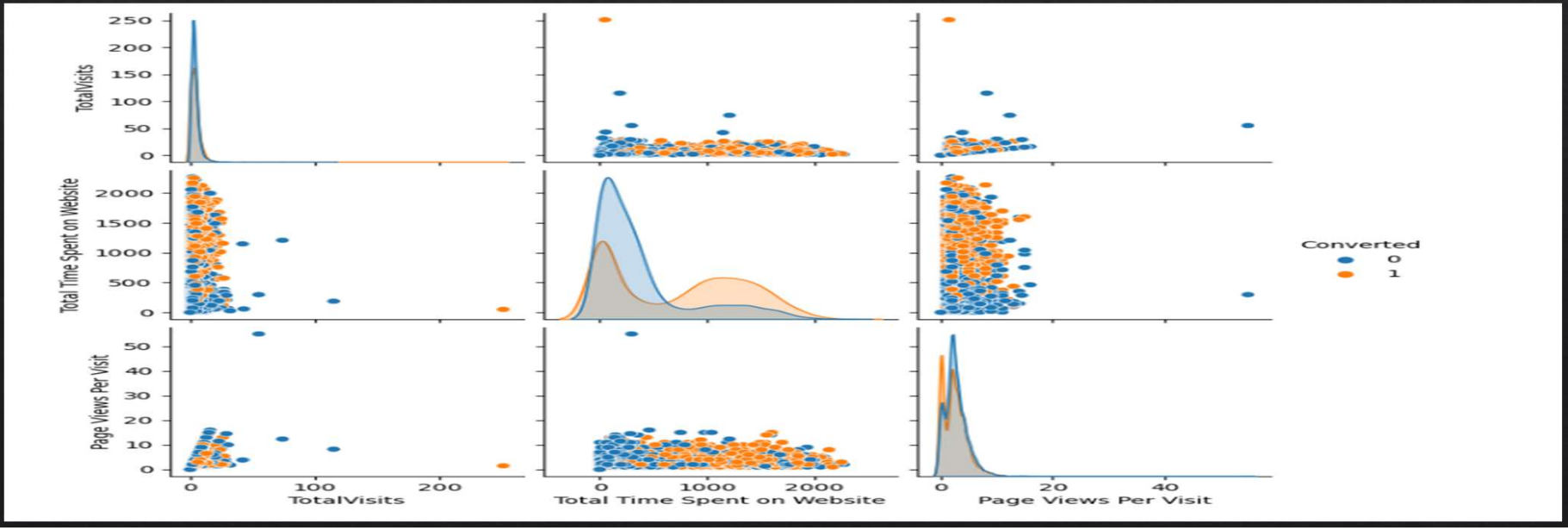
- 'Do Not Call', 'What matters most to you in choosing a course', 'Search', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations' columns dropped due to high data imbalance.

- 'How did you hear about X Education' , 'Lead Profile' , 'Lead Quality' , 'Asymmetric features are dropped because they are having more than 40% of values missing.

- Outliers treatment done of numerical variables.

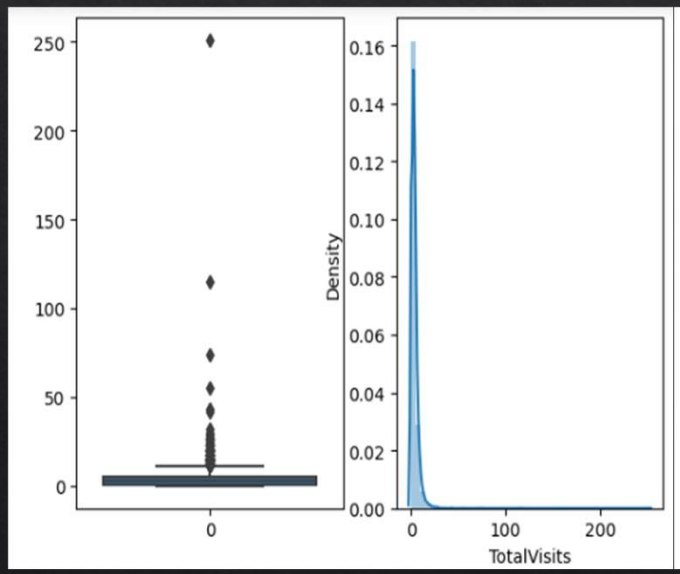
# EDA (Exploratory Data Analysis)

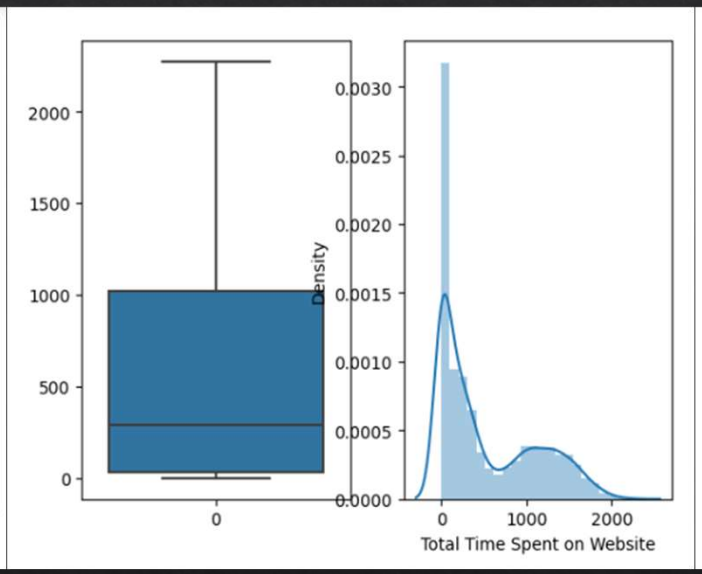
- Univariate Analysis
  - The max probability for TotalVisits is found to be around 15-20 visits. It increases initially but decreases further.
  - The max probability for PageViewsPerVisit is found to be around to be 3-5 pages.
  - The probability of time spent is found to be high for time between 0-300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

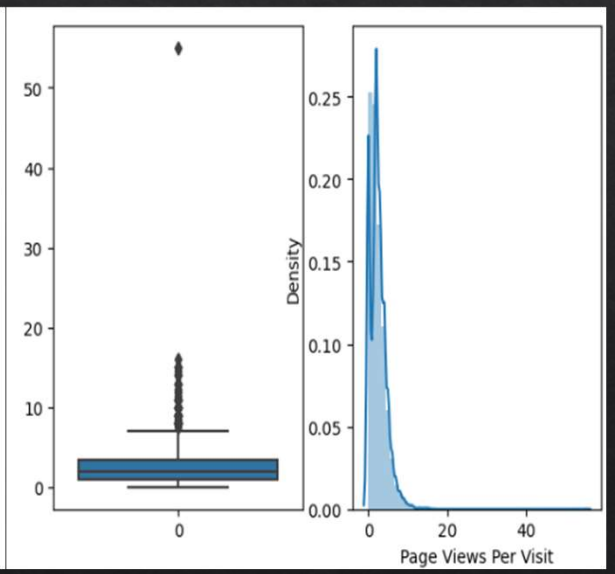


# EDA (Exploratory Data Analysis)

- Univariate Analysis :-
  - 'Total Visits' and 'Page Views per Visit' columns have outliers.



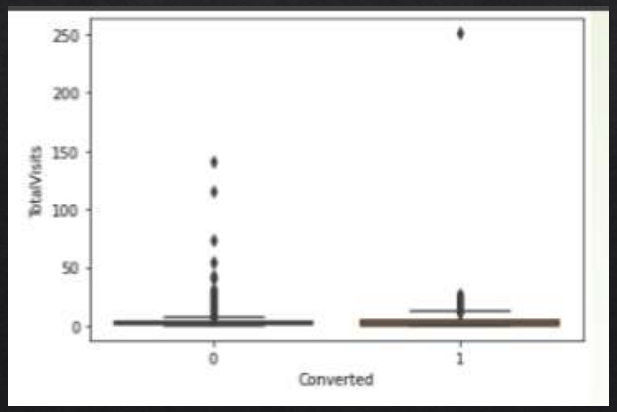



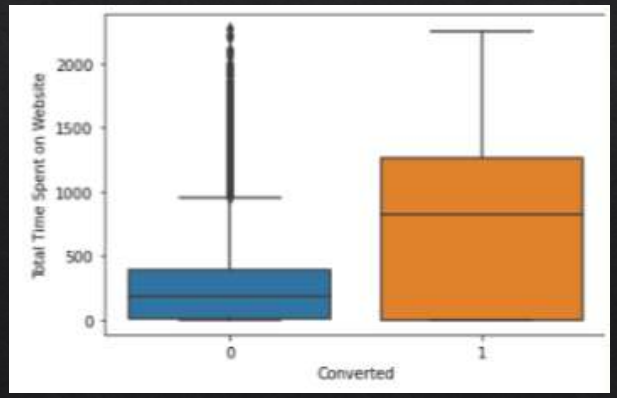



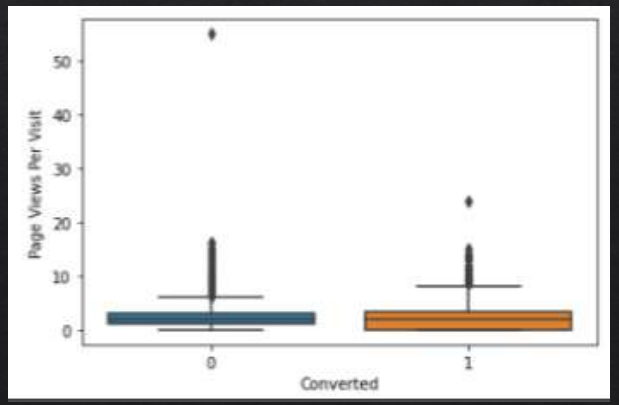
## Bivariate Analysis

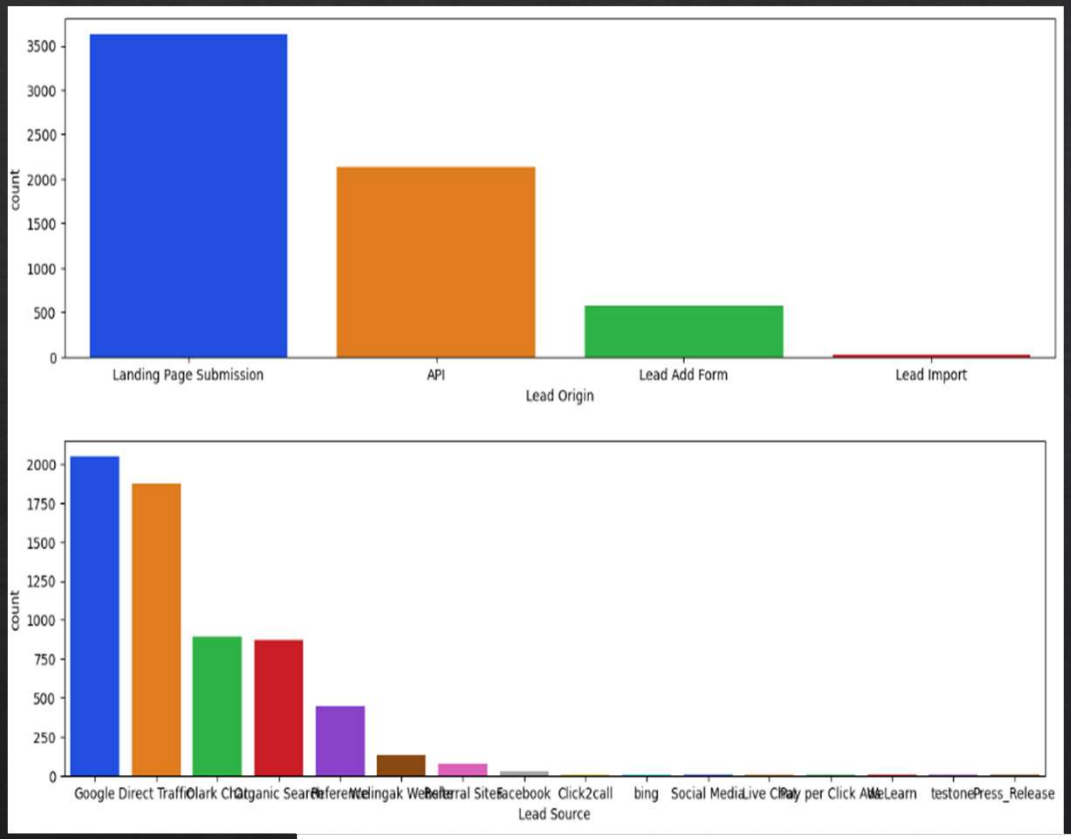
- The mean of time spend on website is found to be higher in case of Converted people rather than non-converted people.
- The average page views for both converted and non converted is found to be the same.
- The average total visits for both converted and non converted people is found to be the same.
- If Lead source is Add Form, the ratio of lead conversion is very high.
- 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.

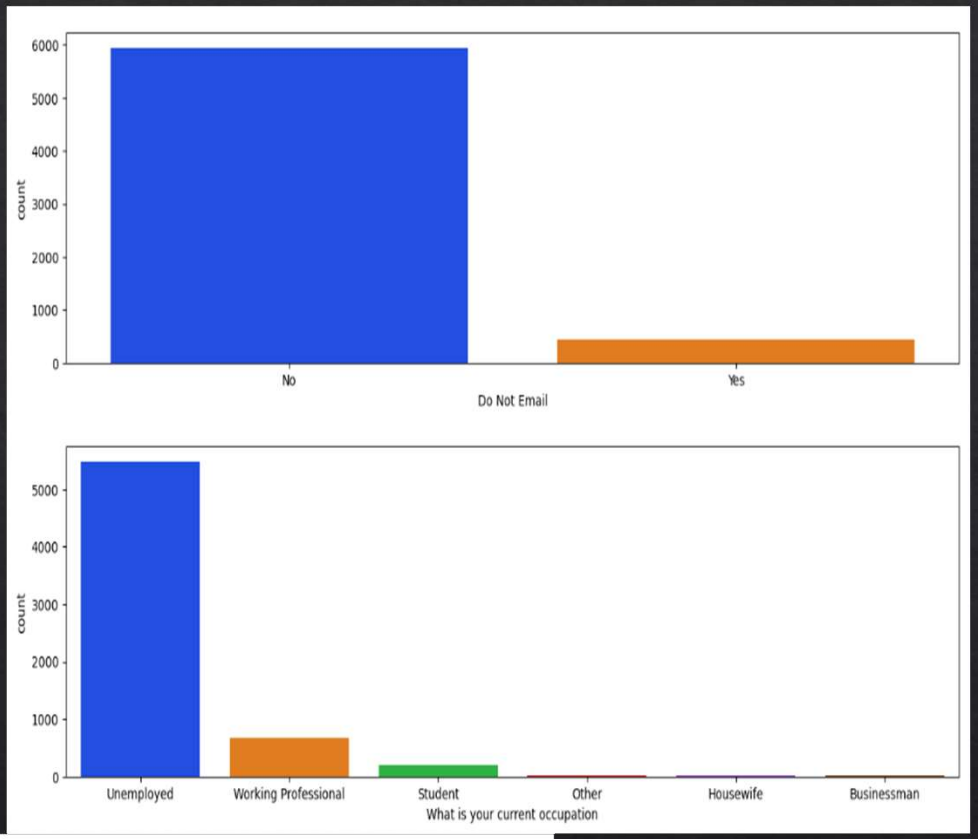
- Google is found to be the important source for Lead Conversion
- It is clearly visible from the graph that we need to target the Unemployed and Working Professional to get a higher conversion rate.
- The ratio of conversion rate is higher than not converted people for working professionals.



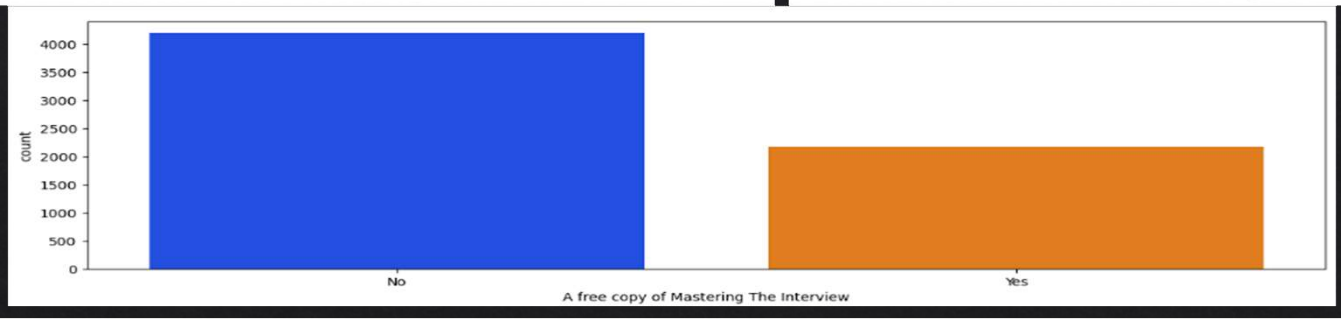




count
Landing Page Submission
API
Lead Add Form
Lead Import
Lead Origin


count
No
Yes
Do Not Email


count
Google
Direct Traffic
Organic Search
Facebook
Click2call
bing
Social Media
testone
Press_Release
Lead Source


count
Unemployed
Working Professional
Student
Other
Housewife
Businessman
What is your current occupation


count
No
Yes
A free copy of Mastering The Interview

## ◈ Observation:-

- Conversion rate for 'API' is close to 'Landing Page Submission'.
- Conversion rate for 'Lead Add Form' number of conversion is more than other.
- So, to improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form and Lead Important.
- Google and Direct traffic generates maximum number of leads.
- Conversion rate of 'Reference' and 'Welingak Website' leads is high.

# CORRELATION MATRIX

From Correlation Matrix, we can see that

- Converted is having positive correlation with Total Time Spent on Website
- and negative relationship with Page Views Per Visit

# Model Building

- Splitting into train and test set.
- Scale variable in train set.
- Build the first model.
- Use RFE to eliminate less relevant variables.
- Build the next Model.
- Eliminate variables based on high p-values.
- Check VIF value for all the existing columns.
- Predict using test set.
- Precision and recall analysis on test predictions.

# Model Building

- Build 1st Logistic Regression training model using all features.
- To build best fit model, we used Recursive Feature Elimination technique to get the top 20 features to build out next model
- For each model build, we have checked for p-value should be less than 0.05
- To remove Multicollinearity, calculated Variance Inflation Factor(VIF) to check if feature variables are not correlated with each other.
- Dropped the features which have high p-value and highly correlated one by one and recursively build the model to get optimal model.

# Final Model Features (VIFs and P- values)

| | Features | VIF |
|---|---|---|
| 0 | Total Time Spent on Website | 1.94 |
| 12 | Last Notable Activity_Modified | 1.83 |
| 2 | Lead Source_Direct Traffic | 1.76 |
| 1 | Lead Origin_Lead Add Form | 1.62 |
| 10 | Last Activity_SMS Sent | 1.56 |
| 4 | Lead Source_Welingak Website | 1.35 |
| 18 | Specialization_Finance Management | 1.33 |
| 6 | Last Activity_Converted to Lead | 1.30 |
| 22 | Specialization_Marketing Management | 1.28 |
| 20 | Specialization_Human Resource Management | 1.28 |
| 3 | Lead Source_Organic Search | 1.28 |
| 8 | Last Activity_Olark Chat Conversation | 1.26 |
| 11 | What is your current occupation_Working Profes... | 1.24 |
| 9 | Last Activity_Page Visited on Website | 1.17 |
| 23 | Specialization_Operations Management | 1.16 |
| 25 | Specialization_Supply Chain Management | 1.15 |
| 15 | Specialization_Business Administration | 1.14 |
| 21 | Specialization_IT Projects Management | 1.13 |
| 26 | Specialization_Travel and Tourism | 1.11 |
| 14 | Specialization_Banking, Investment And Insurance | 1.11 |
| 5 | Do Not Email_Yes | 1.11 |
| 19 | Specialization_Healthcare Management | 1.06 |
| 17 | Specialization_E-COMMERCE | 1.04 |
| 24 | Specialization_Rural and Agribusiness | 1.03 |
| 16 | Specialization_E-Business | 1.02 |
| 13 | Last Notable Activity_Unreachable | 1.02 |
| 7 | Last Activity_Had a Phone Conversation | 1.01 |

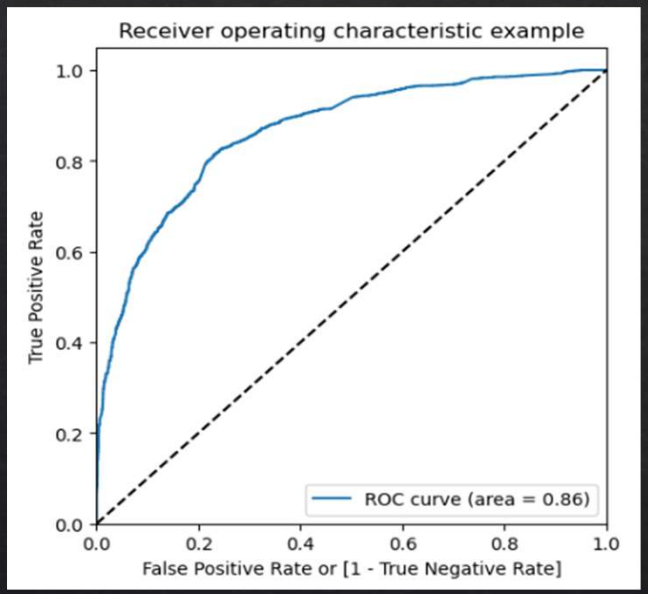
Generalized Linear Model Regression Results

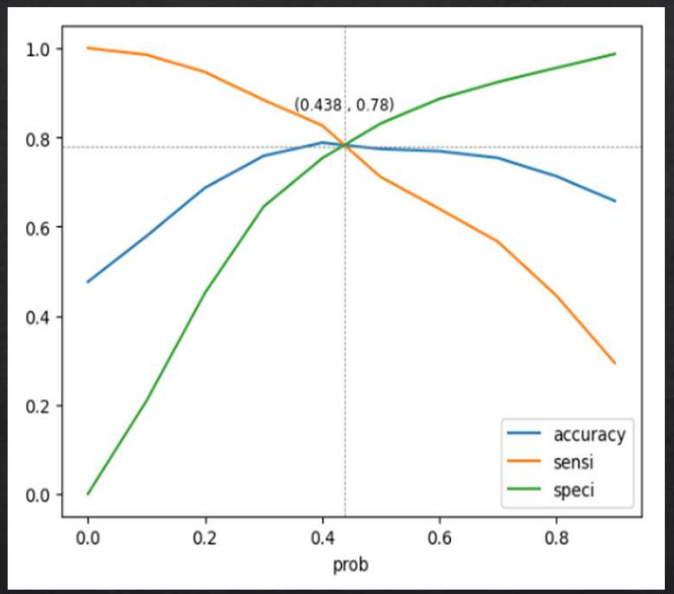
| | | | |
|---|---|---|---|
| **Dep. Variable:** | Converted | **No. Observations:** | 4392 |
| **Model:** | GLM | **Df Residuals:** | 4363 |
| **Model Family:** | Binomial | **Df Model:** | 28 |
| **Link Function:** | Logit | **Scale:** | 1.0000 |
| **Method:** | IRLS | **Log-Likelihood:** | -1979.1 |
| **Date:** | Mon, 20 Nov 2023 | **Deviance:** | 3958.2 |
| **Time:** | 20:38:52 | **Pearson chi2:** | 4.53e+03 |
| **No. Iterations:** | 7 | **Pseudo R-squ. (CS):** | 0.3829 |
| **Covariance Type:** | nonrobust | | |

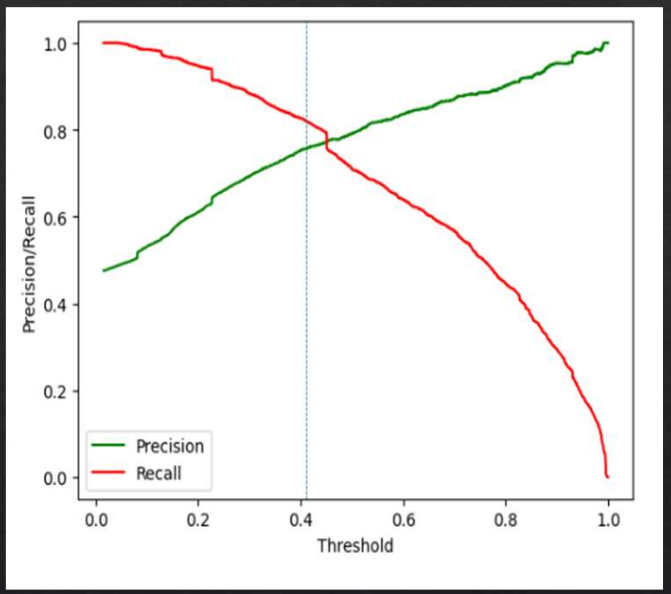
| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -1.0956 | 0.093 | -11.782 | 0.000 | -1.278 | -0.913 |
| Total Time Spent on Website | 3.9928 | 0.178 | 22.400 | 0.000 | 3.643 | 4.342 |
| Lead Origin_Landing Page Submission | -1.1234 | 0.124 | -9.061 | 0.000 | -1.366 | -0.880 |
| Lead Origin_Lead Add Form | 2.3331 | 0.221 | 10.543 | 0.000 | 1.899 | 2.767 |
| Lead Source_Direct Traffic | -0.4087 | 0.105 | -3.895 | 0.000 | -0.614 | -0.203 |
| Lead Source_Organic Search | -0.3815 | 0.124 | -3.083 | 0.002 | -0.624 | -0.139 |
| Lead Source_Welingak Website | 2.9905 | 1.031 | 2.901 | 0.004 | 0.970 | 5.011 |
| Do Not Email_Yes | -1.3429 | 0.197 | -6.825 | 0.000 | -1.729 | -0.957 |
| Last Activity_Converted to Lead | -0.6570 | 0.237 | -2.777 | 0.005 | -1.121 | -0.193 |
| Last Activity_Had a Phone Conversation | 2.3281 | 0.953 | 2.444 | 0.015 | 0.461 | 4.195 |
| Last Activity_Olark Chat Conversation | -0.7175 | 0.199 | -3.598 | 0.000 | -1.108 | -0.327 |
| Last Activity_Page Visited on Website | -0.4243 | 0.177 | -2.394 | 0.017 | -0.772 | -0.077 |
| Last Activity_SMS Sent | 1.0559 | 0.088 | 12.007 | 0.000 | 0.884 | 1.228 |
| What is your current occupation_Working Professional | 2.4481 | 0.196 | 12.480 | 0.000 | 2.064 | 2.833 |
| Last Notable Activity_Modified | -0.6997 | 0.102 | -6.883 | 0.000 | -0.899 | -0.500 |
| Last Notable Activity_Unreachable | 2.2092 | 0.816 | 2.709 | 0.007 | 0.611 | 3.808 |
| Specialization_Banking, Investment And Insurance | 1.0039 | 0.220 | 4.566 | 0.000 | 0.573 | 1.435 |
| Specialization_Business Administration | 0.7724 | 0.200 | 3.866 | 0.000 | 0.381 | 1.164 |
| Specialization_E-Business | 0.8729 | 0.500 | 1.745 | 0.081 | -0.108 | 1.854 |
| Specialization_E-COMMERCE | 0.6915 | 0.364 | 1.901 | 0.057 | -0.021 | 1.404 |
| Specialization_Finance Management | 0.6745 | 0.152 | 4.446 | 0.000 | 0.377 | 0.972 |
| Specialization_Healthcare Management | 0.8369 | 0.296 | 2.829 | 0.005 | 0.257 | 1.417 |
| Specialization_Human Resource Management | 0.7351 | 0.156 | 4.722 | 0.000 | 0.430 | 1.040 |
| Specialization_IT Projects Management | 0.7086 | 0.223 | 3.177 | 0.001 | 0.271 | 1.146 |
| Specialization_Marketing Management | 0.6496 | 0.153 | 4.237 | 0.000 | 0.349 | 0.950 |
| Specialization_Operations Management | 0.8710 | 0.192 | 4.538 | 0.000 | 0.495 | 1.247 |
| Specialization_Rural and Agribusiness | 1.6800 | 0.428 | 3.929 | 0.000 | 0.842 | 2.518 |
| Specialization_Supply Chain Management | 0.7060 | 0.211 | 3.348 | 0.001 | 0.293 | 1.119 |
| Specialization_Travel and Tourism | 0.7182 | 0.264 | 2.718 | 0.007 | 0.200 | 1.236 |

# MODEL EVALUATION – TRAIN DATA

- After getting optimal model, evaluated performance metrics score Accuracy, Recall, Precision, F1 score.
- ROC curve plotted that shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in
- specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.
- Calculated Optimal cutoff point between sensitivity & specificity. From below plot, we have received 0.33 as the optimal cut-off point.
- Also checked Precision and Recall trade-off as this will help us to identify the predicted CONVERTED is actual CONVERTED
- The Precision and Recall tradeoff came out to be 0.38, we have considered that as our cut-off probability on test data.


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.86)
0.0
0.2
0.4
0.6
0.8
1.0


(0.438 , 0.78)
accuracy
sensi
speci
prob
0.0
0.2
0.4
0.6
0.8
1.0


Precision/Recall
Threshold
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0

# MODEL EVALUATION – TEST DATA

- Run the final optimal model on test dataset with below observations
  - ROC Curve came out similar to what we got on our train data.
  - Recall/Sensitivity Score is 81 %
  - Accuracy – 78 %
  - Precision – 76 %

# LEAD SCORE PREDICTION

- The final_predicted column shows the conversion probability of prospective lead

- Lead Score above 39 have a high tendency of converting to a Hot Lead category

# CONCLUSIONS/FINAL INSIGHTS

- For Lead Conversion Rate to be high(more than 80% as per problem statement) we need sensitivity/Recall/TPR to be above 80%.
- If we want to further increase sensitivity we can decrease the cut off by 0.010.02.
- But we cannot decrease further because specificity will decrease.
- There is not much difference between train and test data's performance metrics.
- This implies that our final model didn't over fit training data and is performing well.
- From our model, We can conclude following points :
  1. The customer/leads who fills the form are the potential leads.(Add Form)
  2. We must majorly focus on working professionals.
  3. It’s always good to focus on customers, who have spent significant time on our website.

- Focusing on Hot Leads will increase the chances of obtaining more value to the business as the number of people we contact less no of people and the conversion rate would be high.

# Thank You