LEAD SCORING CASE STUDY

Soumyajit Mitra

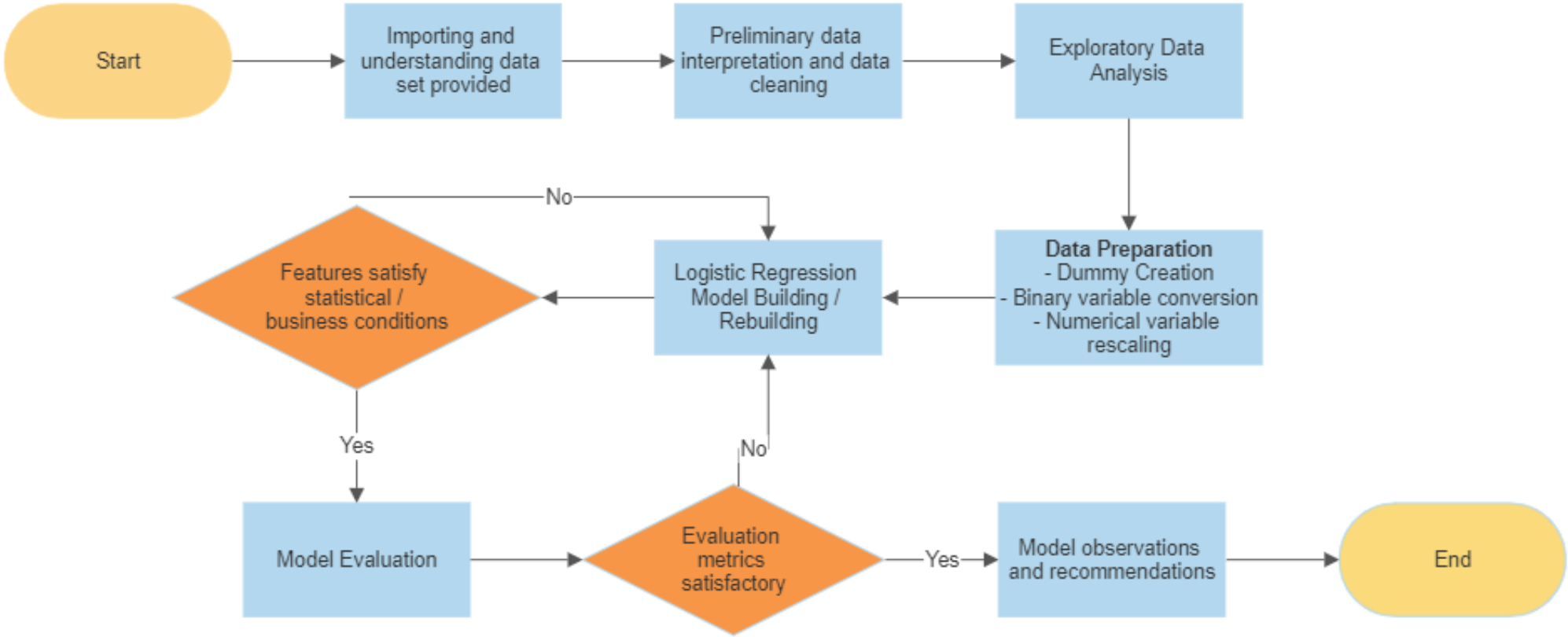
Kalaivannan S

Sonal Verma

PROBLEM STATEMENT

- An education company, X education, sells online courses to industry professionals
- When the interested people visit their websites and fill a form, they become leads for conversion or successful selling of their courses
- Once leads are acquired, the company employees contact the leads through different mediums but the typical lead conversion is 30%, that is, for every 100 people reached out as leads, only 30% people actually buy their courses
- Our model aims to find the most potential leads (hot leads) such that at least 80% of the leads are converted, that is, they buy the courses offered
  - ✓ It serves the purpose of lesser resource expenditure with increased probability of lead conversion.
  - ✓ Chosen Machine Learning technique was Logistic Regression as this was a case of binary classification

APPROACH STRATEGY

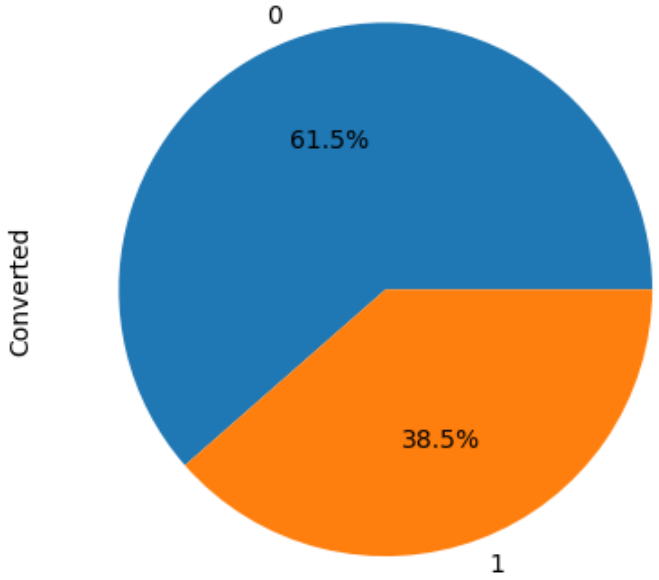


DATA CLEANING

- Removal of columns that had above 30% missing values
- Conversion of ‘Select’ entries in certain columns to null since it signifies that while filling the form, the candidate must not have chosen any category of a particular data field
- Further removal of columns that had above 30% missing values post the above conversion
- Checking the value distributions for each columns, some had faulty entries like “Google” and “google” which convey the same meaning
- Dropping of columns with very high value imbalance (>~99% skew) like ‘Do Not Call’, ‘Magazine’, ‘Search’, ‘Newspaper Article’, ‘Newspaper’ among others since it gives a very less sample size for the model to learn conditions behind a “Yes” in these categories
- Lastly, outlier treatment of numerical variables using medians, or dropping rows

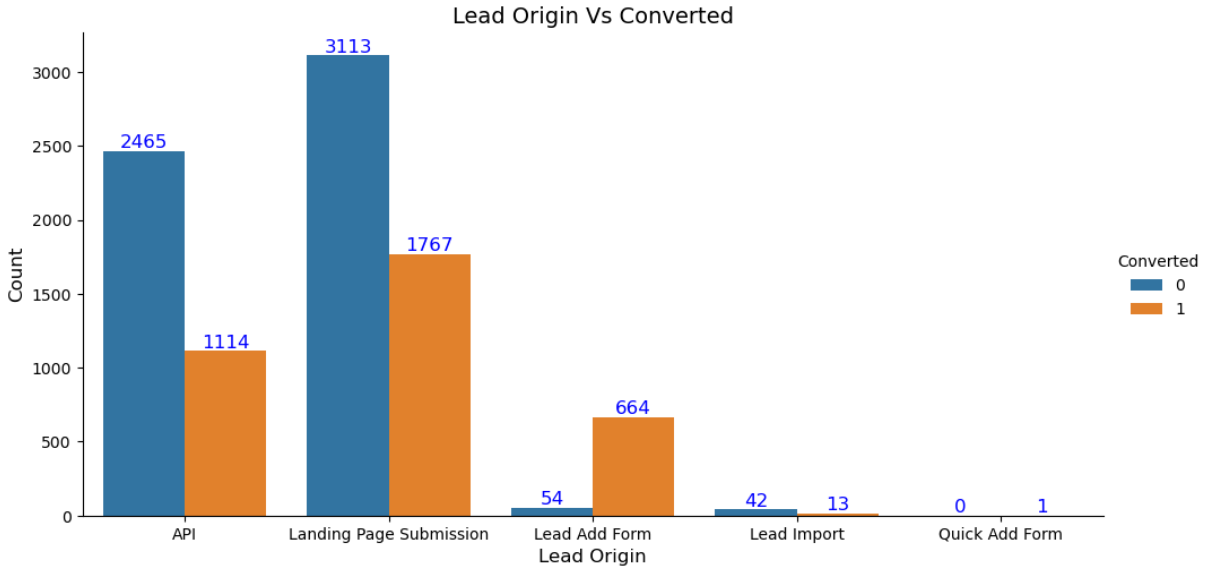
EXPLORATORY DATA ANALYSIS

DATA IMBALANCE



We see a data imbalance of the target variable with 61.5% towards unconverted leads

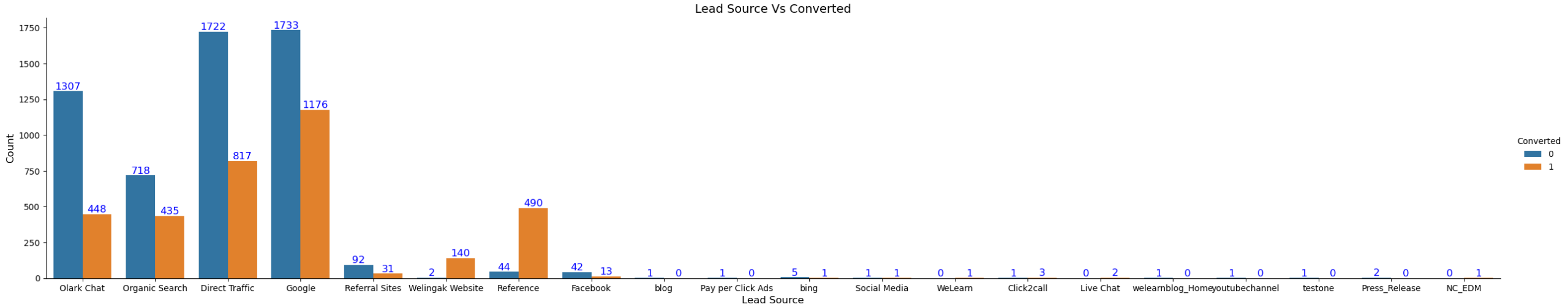
LEAD ORIGIN



Highest conversion rates were observed from Lead Add Form category in the Lead Origin Feature. Rest of the columns either had fewer entries or higher unsuccessful conversion ratio

EXPLORATORY DATA ANALYSIS

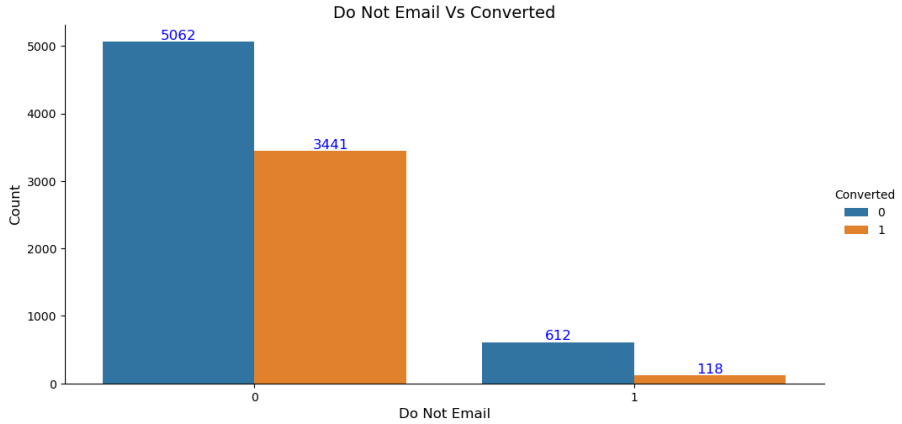
LEAD SOURCE



Highest conversions were from Google whereas the best conversion rates were from Reference, Welingak website, Organic Search and Referral Sites. Rest either had low conversion rates or low count of samples

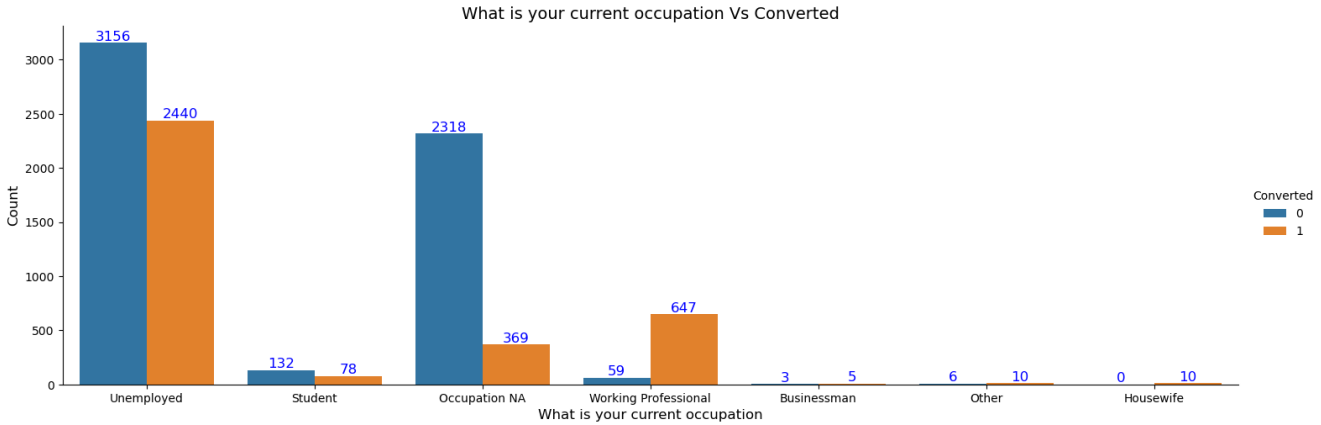
EXPLORATORY DATA ANALYSIS

DO NOT EMAIL



A pretty even distribution for both responses in Do Not Email, while relatively higher conversion rates and conversion total for people opting ‘No’

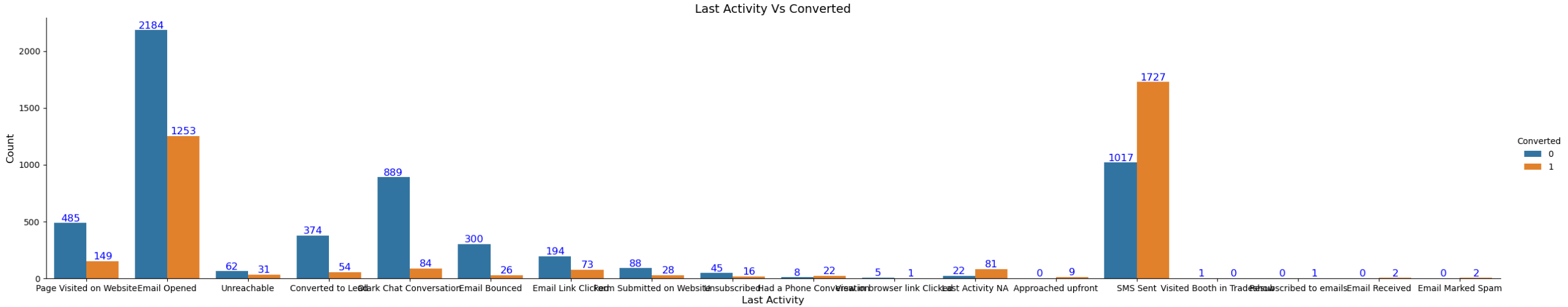
CURRENT OCCUPATION



Highest conversion rates with significant sample size can be seen amongst working professionals and unemployed

EXPLORATORY DATA ANALYSIS

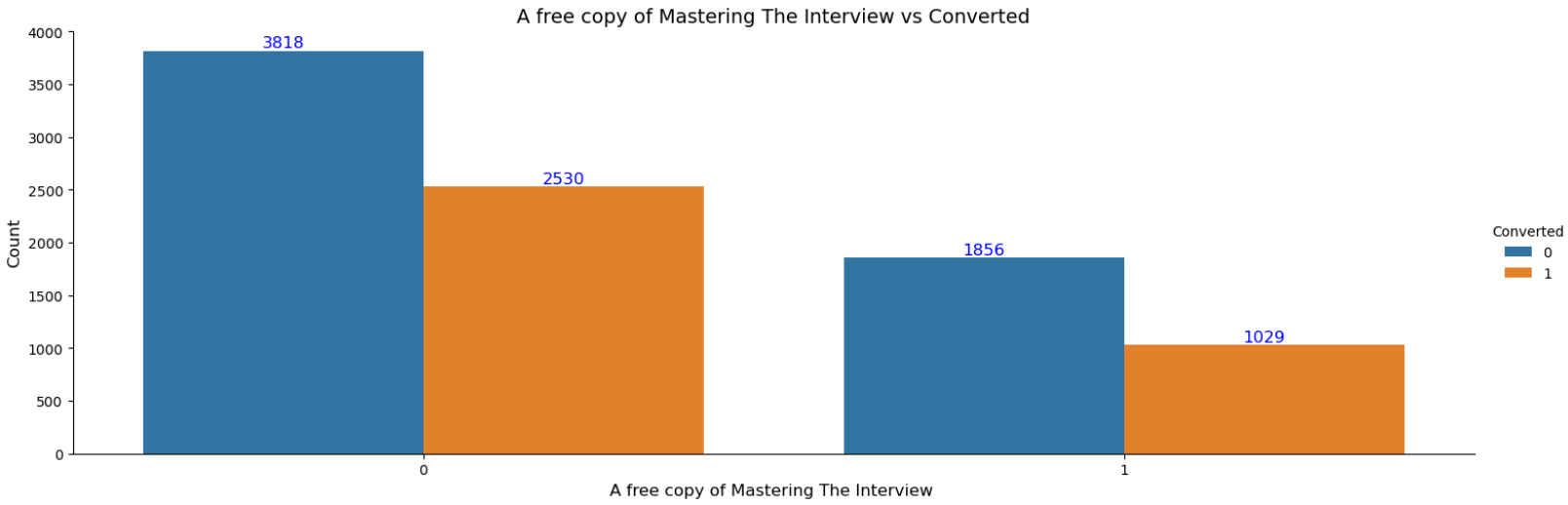
LAST ACTIVITY



The best conversion rates can be seen from SMS Sent and Email Opened categories, also SMS sent has the highest number of conversions, similar trends are observed with Last Notable Activity

EXPLORATORY DATA ANALYSIS

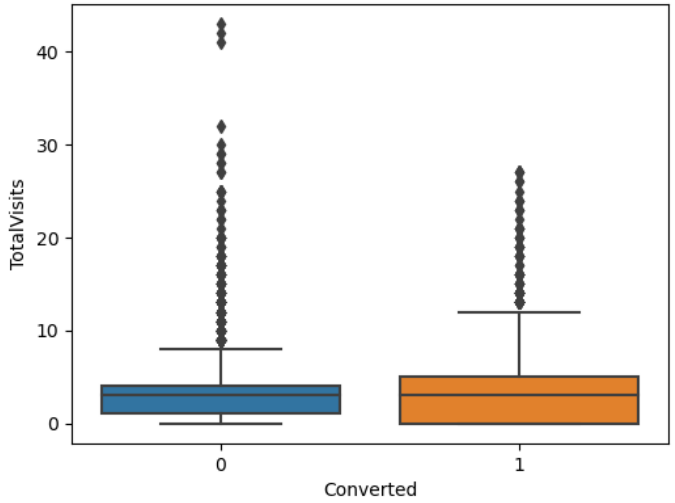
FREE COPY FOR MASTERING THE INTERVIEW

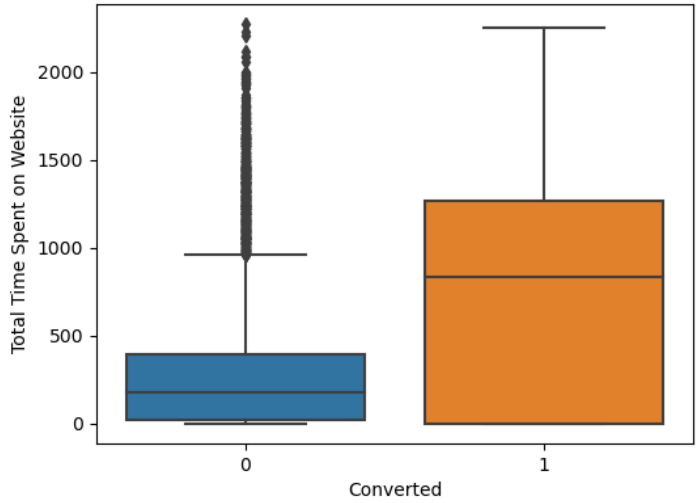


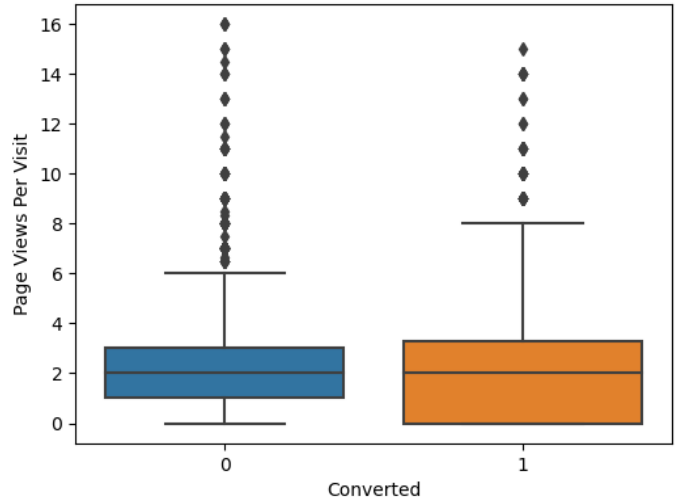
Similar conversion rates observed with both the categories, although people opting for ‘No’ have more number of conversions

EXPLORATORY DATA ANALYSIS

NUMERICAL CATEGORIES







For numerical values it was observed that

- Total visits almost had the same interquartile range and same medians for both converted and non-converted
- Time spent on website for converted people had a much larger range and median than those for not converted, indicating people tend to spend more time on website when interested
- Page views per visit shows that the lower boundary of converted dataset had much less value than those of non-converted, a logical conclusion might be people who are already interested might opt for any course at one go or minimum visits rather than making multiple visits to decide

MODELLING

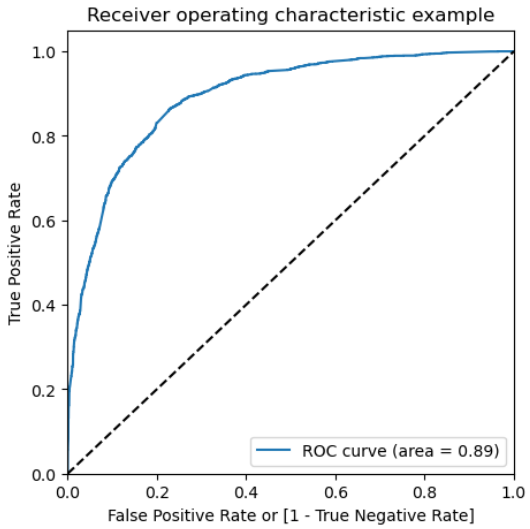
Dep. Variable:	Converted	No. Observations:	6463
Model:	GLM	Df Residuals:	6446
Model Family:	Binomial	Df Model:	16
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2598.7
Date:	Tue, 18 Jul 2023	Deviance:	5197.5
Time:	12:57:34	Pearson chi2:	6.68e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.4110
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.6083	0.117	-13.708	0.000	-1.838	-1.378
Do Not Email	-1.4622	0.170	-8.588	0.000	-1.796	-1.128
TotalVisits	2.4906	0.532	4.678	0.000	1.447	3.534
Total Time Spent on Website	4.5407	0.167	27.176	0.000	4.213	4.868
Page Views Per Visit	-1.4207	0.394	-3.603	0.000	-2.194	-0.648
LeadOrigin_Landing Page Submission	-0.3548	0.091	-3.896	0.000	-0.533	-0.176
LeadOrigin_Lead Add Form	3.1102	0.206	15.079	0.000	2.706	3.514
Lead_Source_Olark Chat	0.9446	0.133	7.084	0.000	0.683	1.206
Lead_Source_Welingak Website	2.6321	1.024	2.570	0.010	0.625	4.640
Last_Activity_Converted to Lead	-0.6339	0.224	-2.834	0.005	-1.072	-0.195
Last_Activity_Olark Chat Conversation	-1.0119	0.171	-5.909	0.000	-1.347	-0.676
Last_Activity_SMS Sent	1.1995	0.076	15.739	0.000	1.050	1.349
Curr_occ_Occupation NA	-1.1921	0.089	-13.469	0.000	-1.366	-1.019
Curr_occ_Working Professional	2.4654	0.182	13.523	0.000	2.108	2.823
Last_N_Activity_Had a Phone Conversation	3.2288	1.155	2.795	0.005	0.965	5.493
Last_N_Activity_Modified	-0.6478	0.085	-7.597	0.000	-0.815	-0.481
Last_N_Activity_Unreachable	1.9445	0.577	3.371	0.001	0.814	3.075

The modelling was done by selection top 20 variables using Recursive Feature Selection and henceforth iterating manually and dropping features with unsuitable p-value and VIFs

A list of final 16 values were obtained as highlighted in the figure on the left

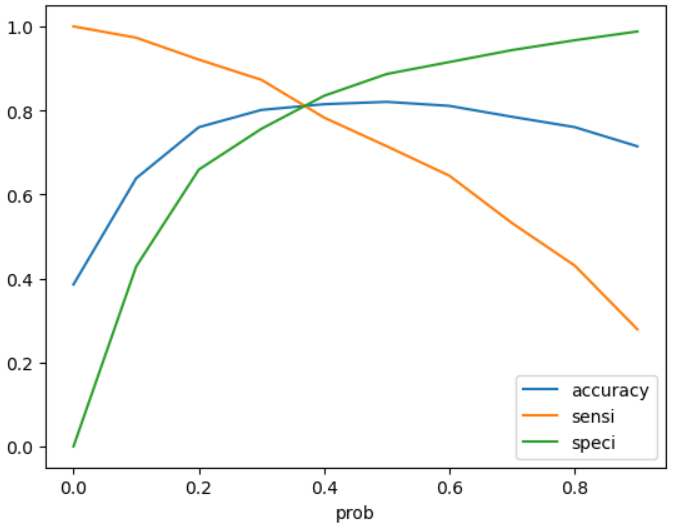
ROC Curve



The ROC curve had an area of 0.89 depicting significant predictive power of the finalised model, hence verifying the model

MODELLING EVALUATION

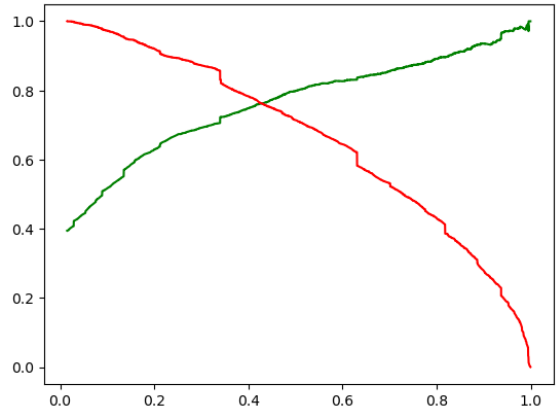
OPTIMAL CUTOFF POINT



From the Accuracy, Sensitivity and Specificity vs Probability Graph, a probability of 0.38 or Score of 38 was finalised as a threshold above which, the leads would be labelled as converted else not converted

CONFUSION MATRIX AND RELATED METRICS

3267	704
512	1980



From the confusion matrix, the below results were obtained:

- Sensitivity = 0.795
- Specificity = 0.823
- False Positive Rate = 0.177
- Positive Predictive Value = 0.738
- Negative Predictive Value = 0.865
- Accuracy = 0.811
- Precision = 0.738
- Recall = 0.795

Although, we were able to find the optimum cutoff from the sensitivity, specificity and accuracy curve, but a precision of ~80% could not be reached, which was the main purpose of the machine learning model. Hence we kept on iterating for appropriate probability score cutoff

MODELLING EVALUATION

ITERATED CUTOFF POINT

We saw that a cutoff probability of 50.5 is a good place to have a desired precision of ~80% and also a good recall of ~71% in the training set

Precision was given importance since as per the business problem, the company wanted a conversion rate of 80% from the predicted hot leads

CONFUSION MATRIX AND RELATED METRICS

3530	441
720	1772

From the confusion matrix, the below results were obtained:

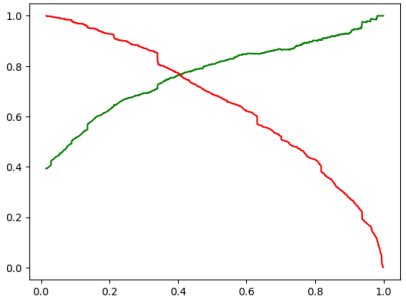
- Sensitivity = 0.711
- Specificity = 0.888
- False Positive Rate = 0.111
- Positive Predictive Value = 0.801
- Negative Predictive Value = 0.830
- Accuracy = 0.820
- Precision = 0.801
- Recall = 0.711

Hence a cutoff score of 50.5 was accepted and applied further to test set

MODELLING EVALUATION (TEST SET)

CONFUSION MATRIX AND RELATED METRICS

1530	173
336	731



From the confusion matrix, the below results were obtained:

- Sensitivity = 0.685
- Specificity = 0.898
- False Positive Rate = 0.102
- Positive Predictive Value = 0.809
- Negative Predictive Value = 0.819
- Accuracy = 0.820
- Precision = 0.809
- Recall = 0.685

Hence it is observed that even on the test set, the precision and positive predictive values hover around 80% which was the goal of the project

FINAL OBSERVATIONS AND RECOMMENDATIONS

- Since, conversion rate among the hot leads were the most important requirement of the business, the main metrics of focus was Precision, Positive Predictive Values and overall accuracy which seemed to satisfy the ballpark value for both the test and train sets, have listed the values below for reference
    - Train set: Precision = 80%; Positive Predictive Value = 80%; Overall Accuracy = 82%; False Positive Rate = 11%
    - Test set : Precision = 81%; Positive Predictive Value = 81%; Overall Accuracy = 82%; False Positive Rate = 10%
- Also False Positive Rate was low ~10% in the test set, which is a good number since the business does not want to spend resources on false leads
- Coming to the model and the best variables, the model was finalised after 6 iterations with 16 variables and an ROC curve area of 0.89, which signifies good predictive power of the model
- **The most significant three variables were:**
    - Total Time Spent on Website: coefficient = 4.5407, explaining that greater time spent on the website might lead to person taking up any course
    - Last Notable Activity Having A Phone Conversation: coefficient = 3.2288, usually interested candidates prefer a phone-call to attain data about courses they are interested in
    - Lead Origin Lead Add Form: coefficient = 3.1102
- The above variables increase the probability of conversion the most in decreasing order from top to bottom, hence recommended to pay more attention to
- **Variables affecting conversion rate the most in a negative way were:**
    - Do Not Email: coefficient = -1.4622, usually uninterested candidates refuse from receiving email notifications
    - Page Views Per Visit: coefficient = -1.4207, interested candidates might visit page once and spend a lot of time before finalizing, but might not visit page multiple times which expresses unsurity
    - Current Occupation Not Available: coefficient = -1.1921, such kinds of people can not be targeted properly with relevant courses or ads, hence, very uncertain whether they will convert
- The above variables decrease the probability of conversion the most in decreasing order from top to bottom, hence recommended to pay less attention to
- It is also recommended to have an overall conversion rate to be a bit lower than 80% from the hot leads, since it is a trade-off between the conversion rate of hot leads and identification of true leads, increasing the precision might lead to not capturing some leads which might actually have been converted