Lead Scoring Case Study

ASWIN JOSEPH

SHIVA KUMAR

Problem Statement

An education company named X Education sells online courses to industry professionals.

The company markets its courses on several websites and search engines like Google. When 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%.

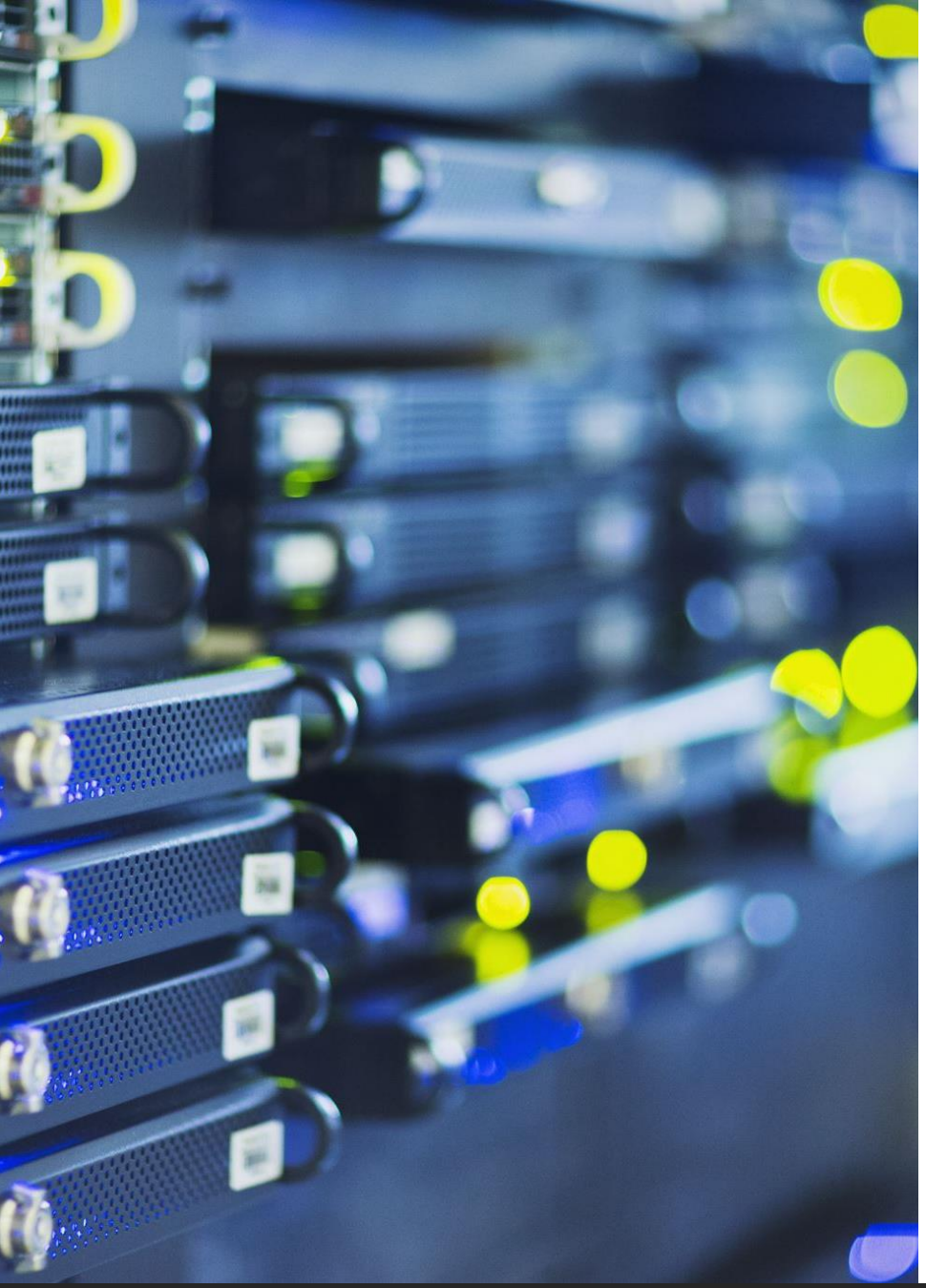
Now, although X Education gets a lot of leads, its lead conversion rate is very poor. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.

Goals of the Case Study

1. 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.
2. 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. They need a suggestion on a good strategy they should employ at this stage.
3. 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. They need a suggestion on a strategy they should employ at this stage also.

Approach

1. **Data Cleaning**
2. **Exploratory Data Analysis**
3. **Assigning Dummy Variables to Categorical variables**
4. **Scaling**
5. **Train-Test Split**
6. **Model Building**
7. **Model Evaluation**
8. **Prediction**



Exploratory Data Analysis

Read the data from CSV file

Outlier Treatment

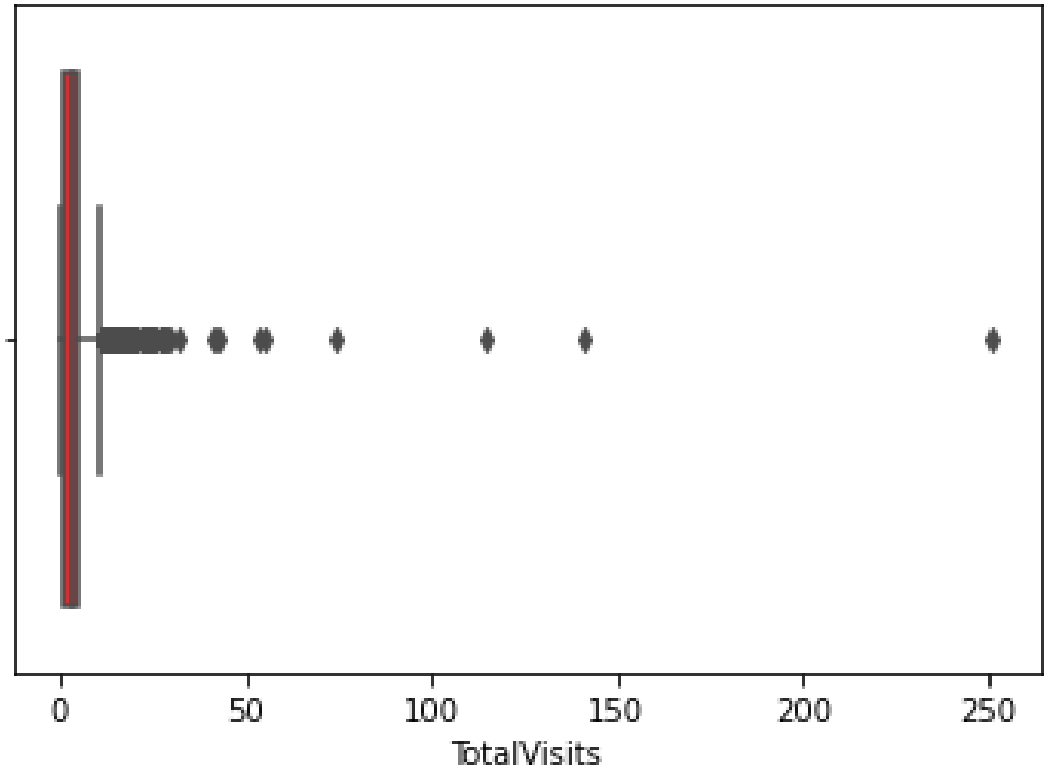
Removing higher null value data – Greater than 45% are removed in our case.

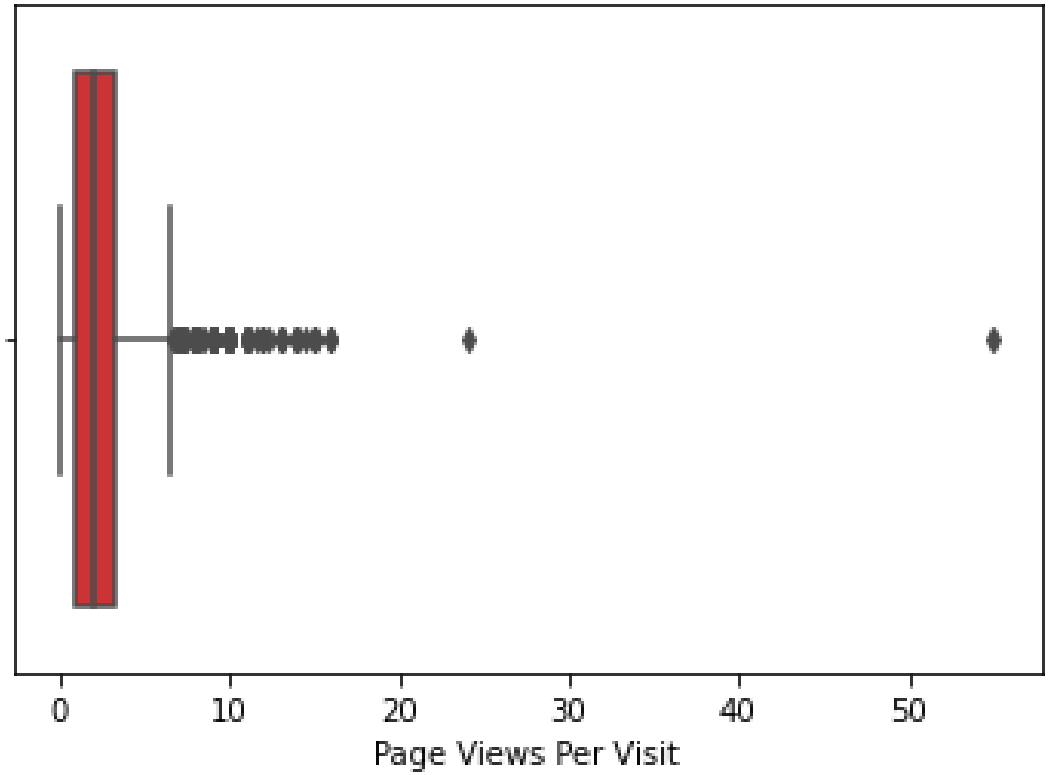
Imputing null values in case where ‘Select’ was found

Feature Standardisation

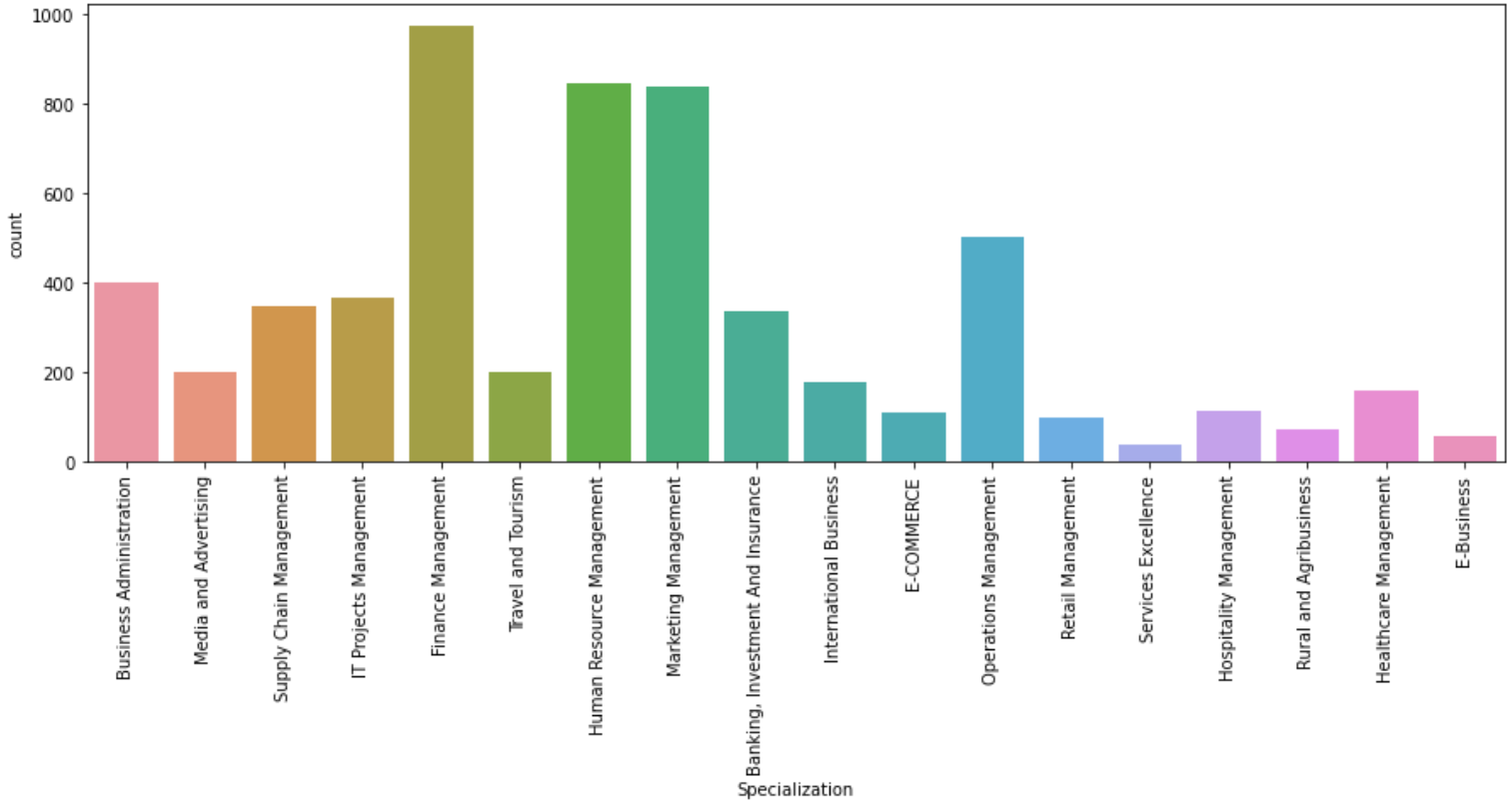
The data frame is now of shape (9074, 14) – before dummy creation from initial (9240, 37)

Outliers are identified and Corrected

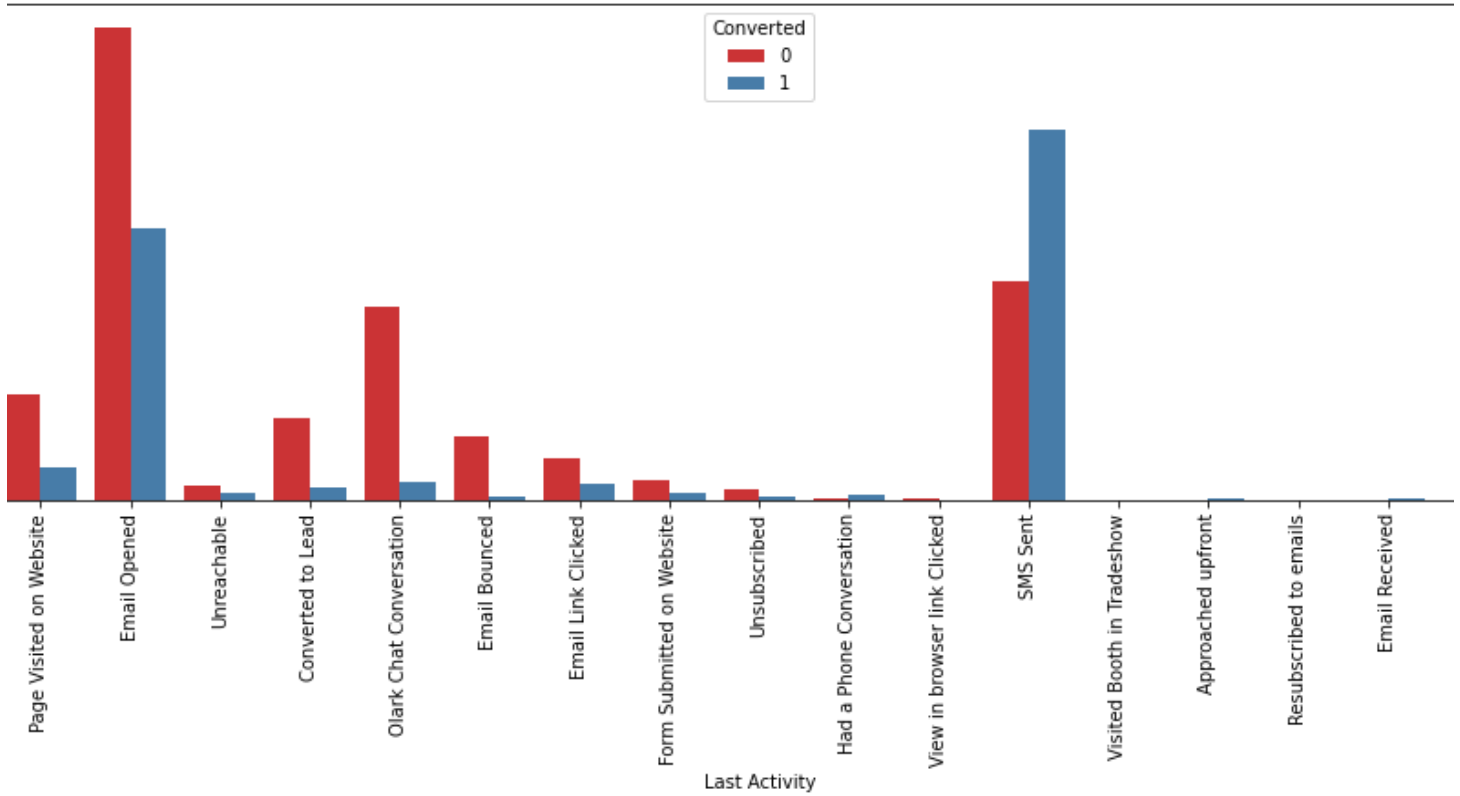




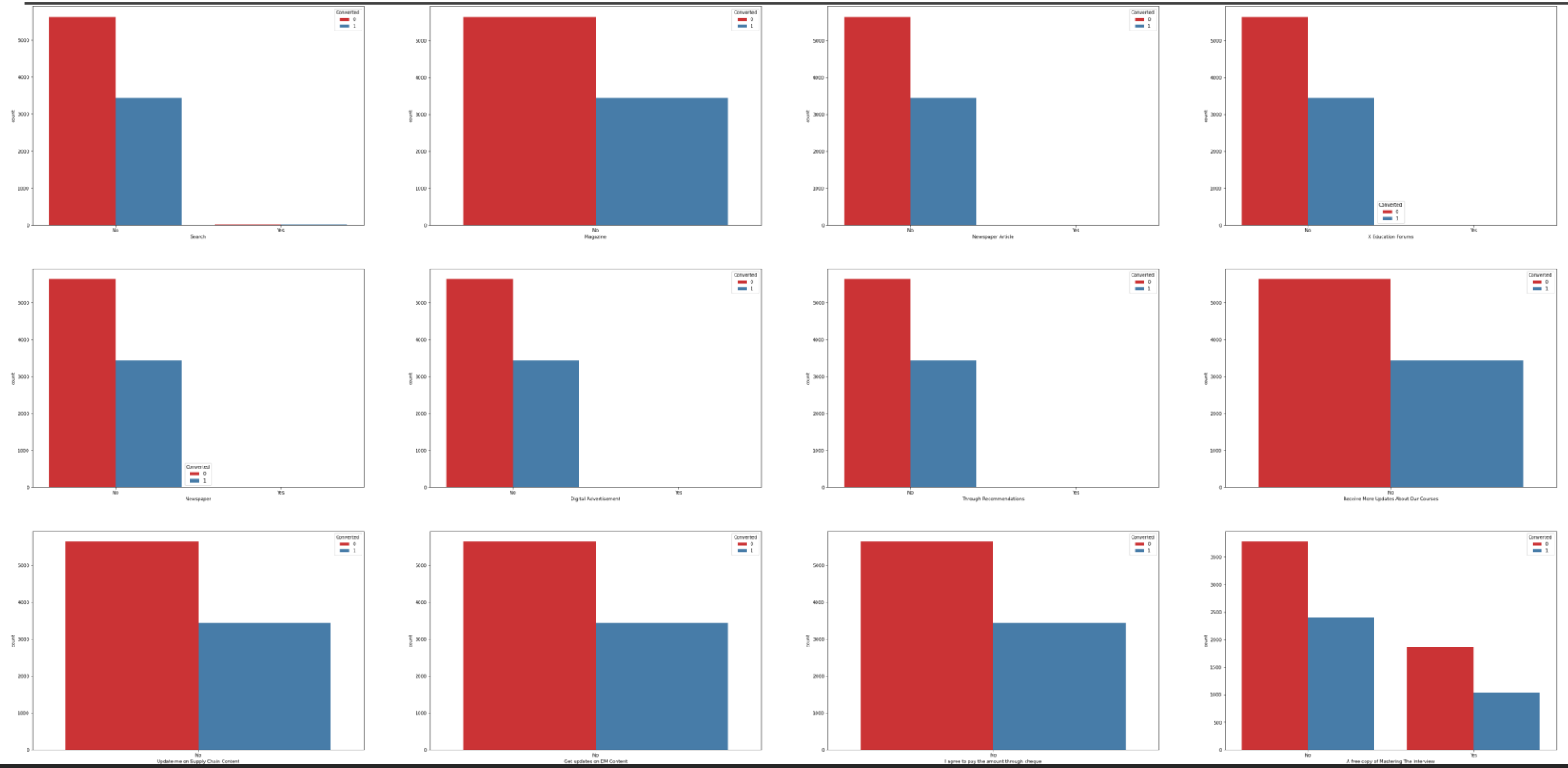
Univariate Analysis - Specialization



Bivariate Analysis – Last Activity



Bivariate Analysis - Features with similar Characteristics with no significant Contribution



Data Preparation & Scaling

Converted Binary variables like YES/No to 1/0

Categorical features were found as "Lead Origin', 'Lead Source', 'Last Activity', 'Specialization', 'What is your current occupation’, 'City’, 'Last Notable Activity’

Dummy variables for categorical features were created

Feature Scaling was done using standard scaler

Model Building

Train and test data were split in 70:30 ratio.

Feature Scaling was done – Top 20 features were selected

9 models were created by dropping features based on significance, P values and VIF

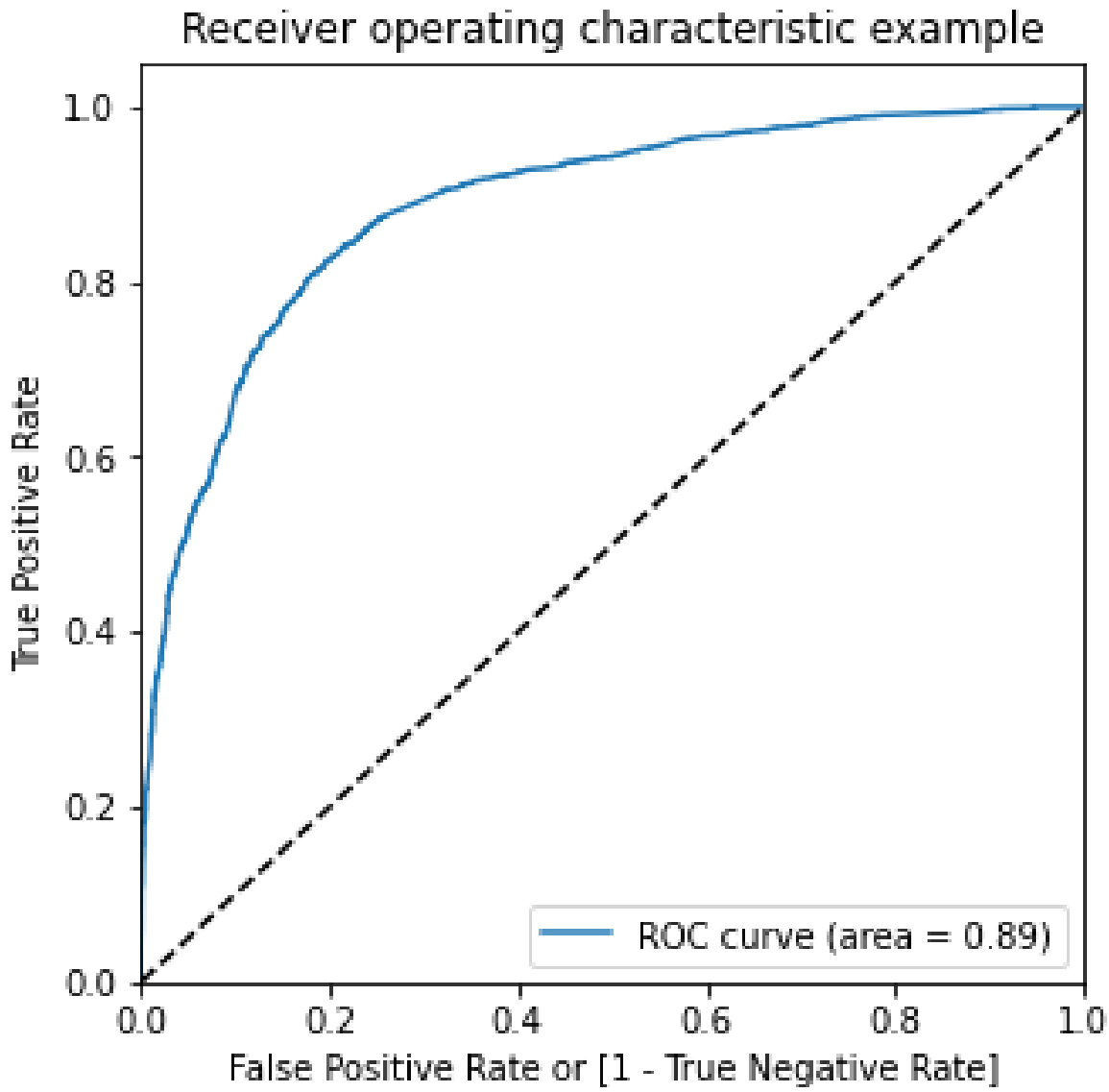
9^{th} model was selected as final model with 12 features

Confusion matrix was created

ROC Curve was plotted and found area under the curve to be 0.89 which depicts a good model.

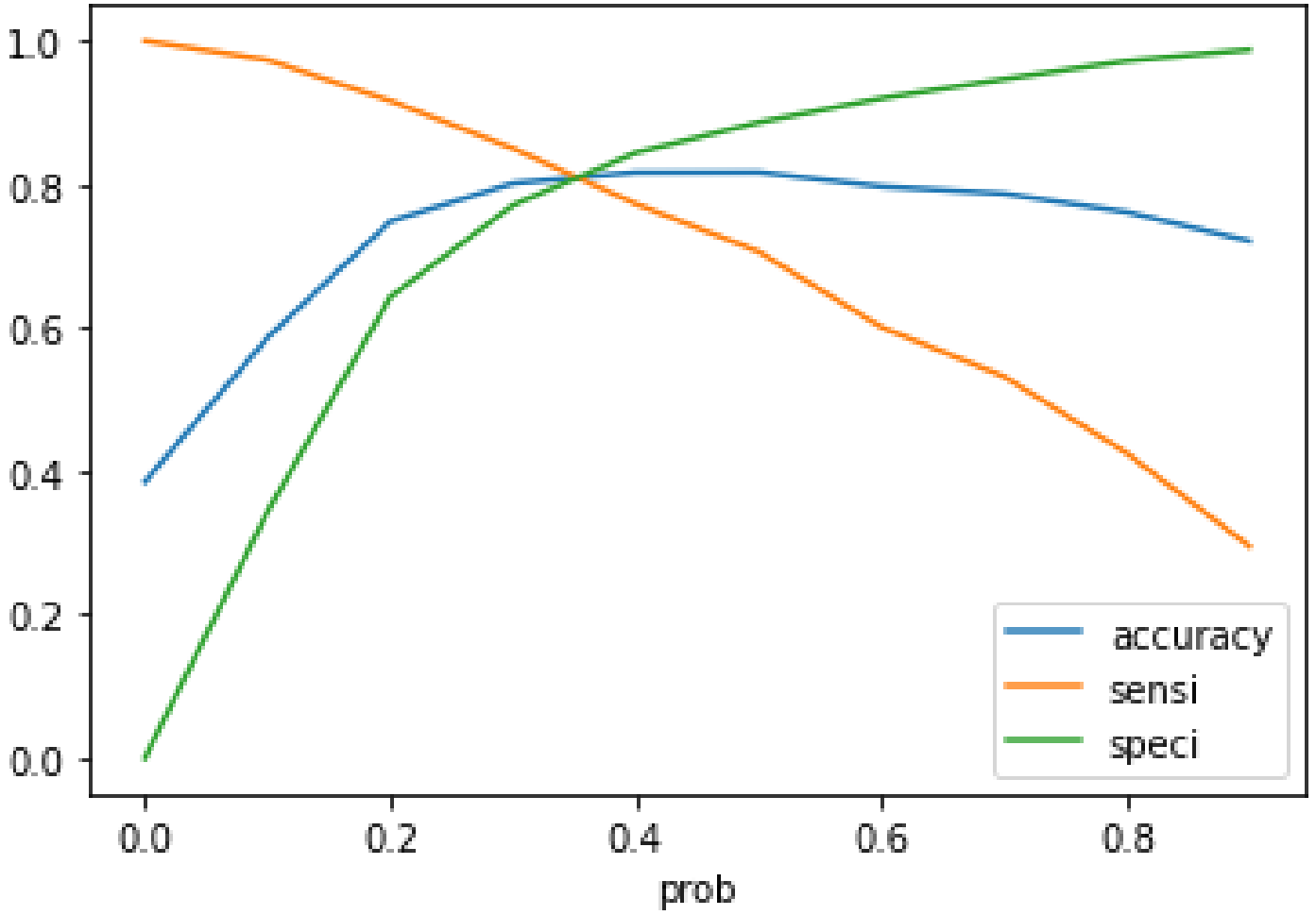
ROC Curve

Area under the curve is 0.89



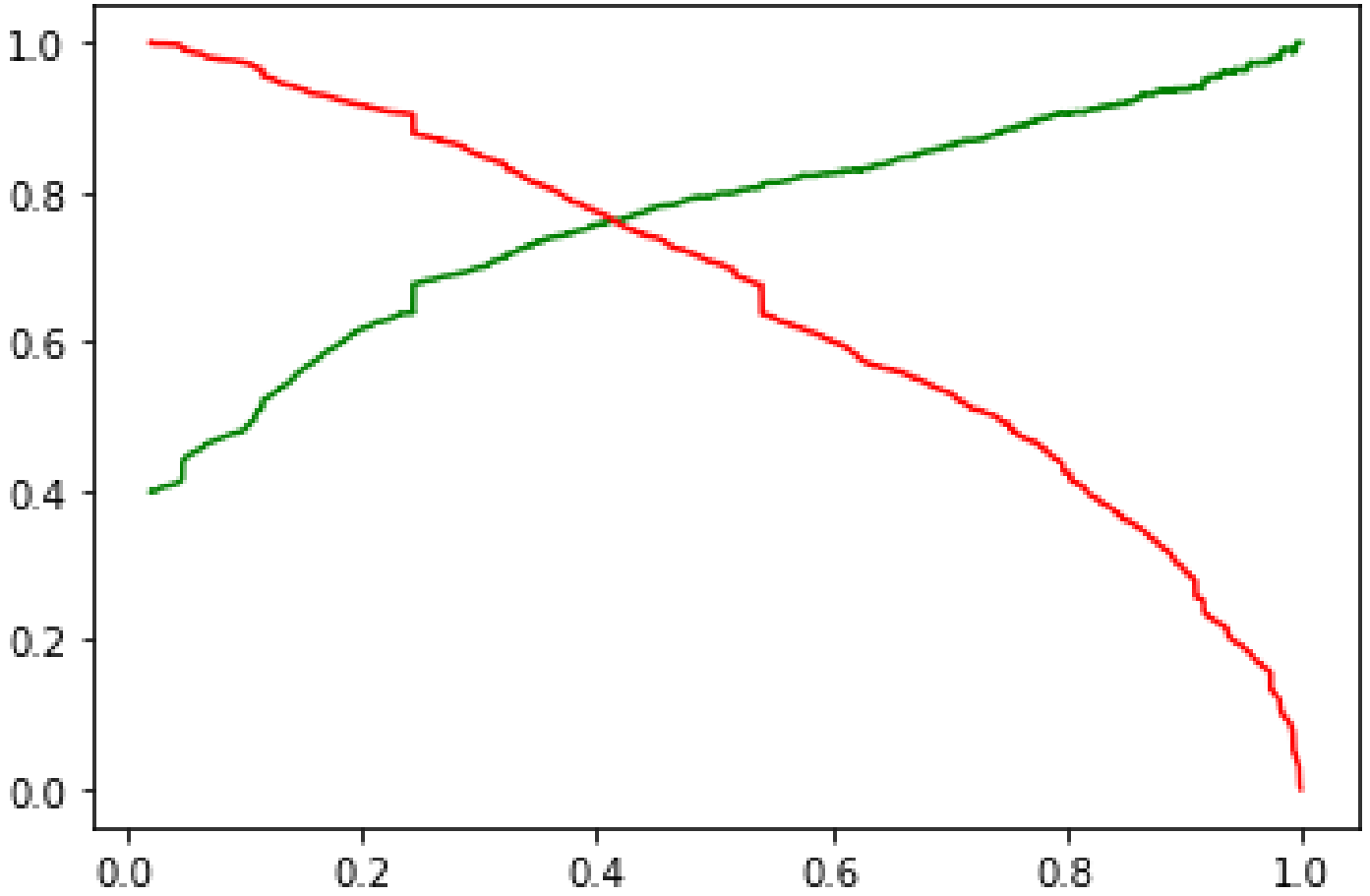
Accuracy, Sensitivity and Specificity

From the curve, 0.34 is found to be the optimum point to take as a cutoff probability.



Trade-off curve

Seems to be matching - Cutoff is coming to nearly 0.41



Evaluation Result

Comparing the values obtained for Train & Test:

Train Data:

Accuracy : 81.0 %, Sensitivity : 81.7 %, Specificity : 80.6 %

Test Data:

Accuracy : 80.4 %, Sensitivity : 80.4 %, Specificity : 80.5 %

Thus, target lead conversion rate using this model is around 80%.

This Model seems to predict the Conversion Rate as desired and decision shall be made in making good calls to get a higher lead conversion rate of 80% from roughly 38% in the raw data.

Final Features – Decreasing order

Lead Source_Welingak Website	5.811465
Lead Source_Reference	3.316598
What is your current occupation_Working Professional	2.608292
Last Activity_Other_Activity	2.175096
Last Activity_SMS Sent	1.294180
Total Time Spent on Website	1.095412
Lead Source_Olark Chat	1.081908
const	-0.037565
Last Notable Activity_Modified	-0.900449
Last Activity_Olark Chat Conversation	-0.961276
Lead Origin_Landing Page Submission	-1.193957
Specialization_Others	-1.202474
Do Not Email	-1.521825

X education’s – 2-month period with 10 interns

During this phase, they shall contact all the leads which have the potential to have positive conversions.

- The customers which are to be contacted can be identified based on "Lead Score" equal to or greater than 85 (in this model). They are termed as 'Hot Leads’.
- These Hot Leads came to be around 368 in numbers who can be straight away contacted.
- If needed to be more aggressive, we can change the “Lead Score” to other values like 80 or even 75 (or whatever decision the management takes), to include more potential Leads and make a call to them.

X education's – Period when focusing on New work

By tweaking the threshold of the lead score (like above 90 or so), the number of Hot leads to be contacted can be fine-tuned.

Also, automated methods like auto SMS and emails shall be configured to be sent to these Hot Leads.

From the final model, we can see that "**Working Professional"** is a significant contributor, hence if a "Working Professional" approaches, they shall be given utmost importance as they are high potential lead which is a clear and simple indication of a Hot Lead.

Thanks