Lead Scoring Case Study

SUBMISSION

Author:
Rajat Gupta
Amit Sood
Prakhar Gupta
V. Sumanth George

Lead Scoring – Case Study Abstract

The purpose of the case study is to identify the most potential leads i.e. ‘Hot Leads’.

Strategy:

We will create a Logistic Regression model and assign a lead score to each lead.

Case Study Approach:

- Delineate between essential and non-essential attributes via data quality process.
- Identify the driving attributes for lead conversion, explain the interdependencies and strategy to employ in various scenarios.

Case Study Analysis Process

Data Extraction and Preparation

- Extract Data and Import them into Python notebook
- Data Cleanup i.e. check for duplicates and null values
- Remove non-essential columns and Impute necessary columns.

Exploratory Data Analysis

- Map Yes/No columns to 0/1, replace Select values with NaN
- Dropping columns which have high % of null
- Perform outlier analysis and remove outliers
- Create Correlation plot for Numerical variables

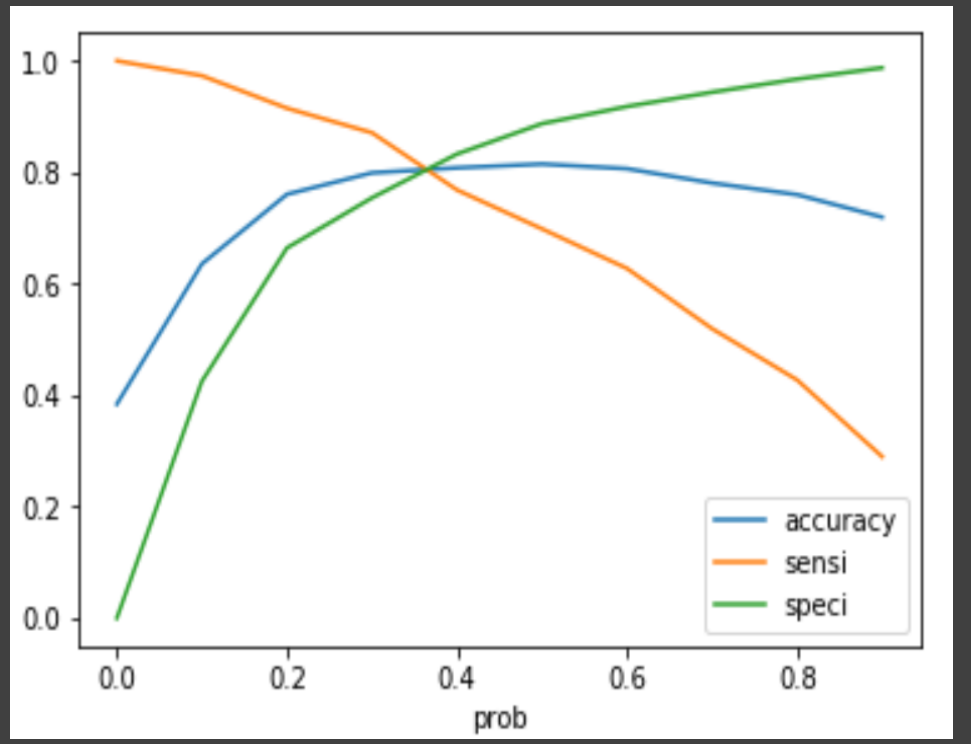
Model Building

- Create Dummy variables and split data in train & test sets.
- Perform standardization of numerical variables
- Implement Sklearn logistic model and use RFE for feature selection
- Assess the features selected by RFE using StatsModels
- Remove columns which have P>0.05 and VIF>5

Model Evaluation

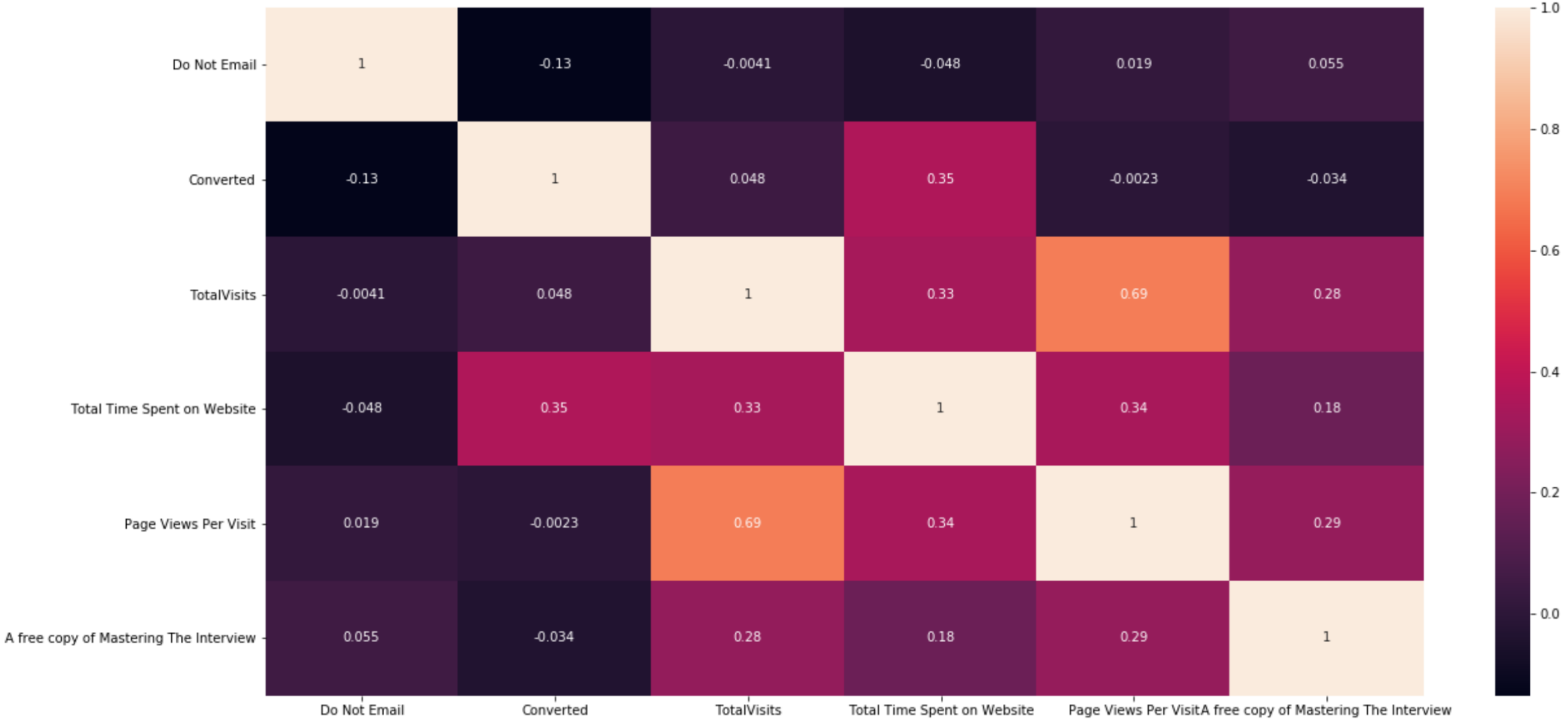
- Find optimal cutoff point for accuracy
- Generate final model with lead number, probability, Lead Score
- Validate the Evaluation Parameters i.e. Accuracy, Sensitivity, Specificity, Precision
- Check the ROC Curve
- Apply prediction on test data

Analysis & Model Building



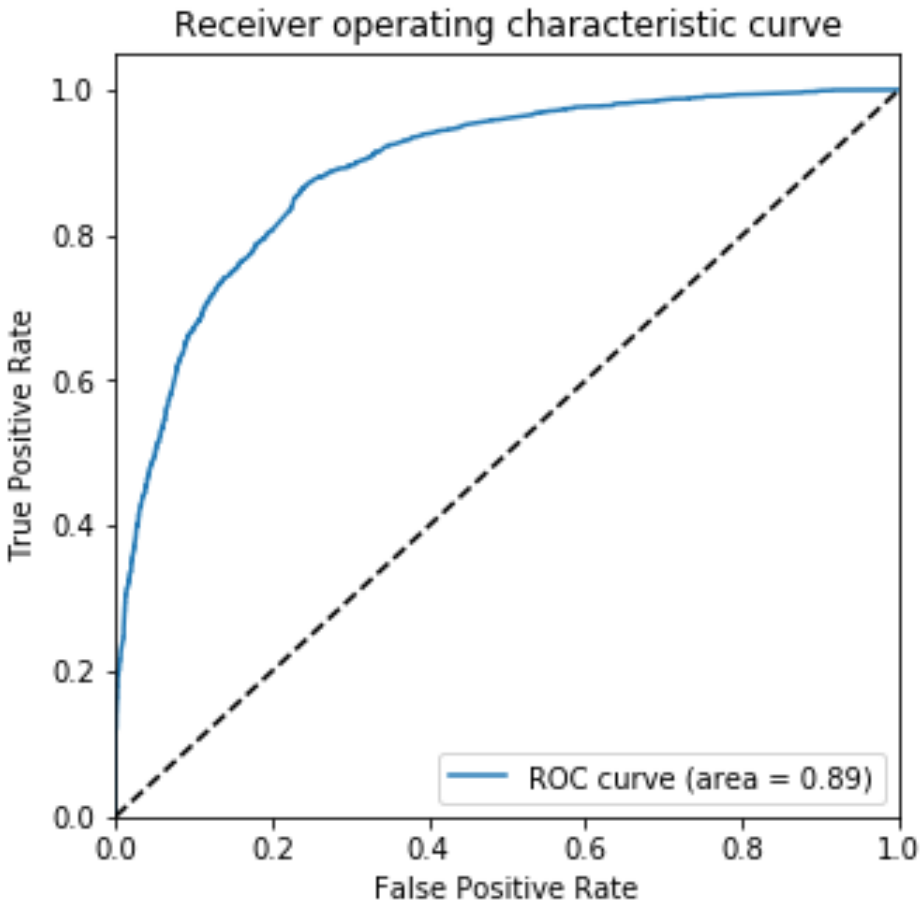
- After Data Clean-up and doing necessary imputations, we did Outlier Analysis and found that in few variables there is sudden increase in value after 95% quantile although they are less in number. Hence, we removed the values from TotalVisits, Total Time Spent on Website, Page Views Per Visit variables which are not in 25% to 95% range.
- On plotting correlation graph between Numerical variables, we observed that there is high correlation in TotalVisits and Page Views Per Visit variables.
- We used Recursive Feature Elimination(RFE) to select 20 features and assessed model using statsmodel.api to check if the selected features are significant(P-value < 0.05) and have variance in control region i.e. VIF < 5.
- We created different models until all the variables used in model are significant and VIF is in control.
- We calculated Accuracy, Sensitivity, Specificity values at different cutoff values ranging from 0.1 to 0.9 and by plotting curve identified optimal cut-off value of 0.38.

Correlation Matrix



High correlation in Total visits and Page views per visit

Model Evaluation



- Considering Optimal cut-off of 0.38, we assigned the Predicted Conversion value to leads i.e. 0 or 1. We have also added the Lead Score to leads based on the Predicted Lead Probability.
- We created the Confusion Matrix using the Actual and Predicted Conversion values.
- We calculated the accuracy of our model and using the Confusion Matrix, we checked Sensitivity, Specificity, Precision, Recall parameters.
- We also plotted the ROC curve to check the trade-off between True Positive Rate and False Positive Rate.
- Finally, we used our model for doing prediction for our Test data.

Final Model Results

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.3313	0.155	-2.135	0.033	-0.635	-0.027
Do Not Email	-1.0716	0.171	-6.251	0.000	-1.408	-0.736
Total Time Spent on Website	1.0832	0.041	26.611	0.000	1.003	1.163
Lead Origin_Landing Page Submission	-1.0289	0.129	-7.975	0.000	-1.282	-0.776
Lead Origin_Lead Add Form	3.1058	0.238	13.042	0.000	2.639	3.573
Lead Source_Olark Chat	1.1312	0.122	9.272	0.000	0.892	1.370
Lead Source_Welingak Website	2.7713	1.035	2.678	0.007	0.743	4.800
Last Activity_Email Opened	0.4799	0.109	4.395	0.000	0.266	0.694
Last Activity_Had a Phone Conversation	2.8542	0.713	4.002	0.000	1.456	4.252
Last Activity_SMS Sent	1.6084	0.109	14.702	0.000	1.394	1.823
Specialization_unknown	-0.9430	0.126	-7.485	0.000	-1.190	-0.696
What is your current occupation_Other	-1.1727	0.088	-13.354	0.000	-1.345	-1.001
What is your current occupation_Working Professional	2.3228	0.189	12.259	0.000	1.951	2.694
Last Notable Activity_Modified	-0.7681	0.090	-8.514	0.000	-0.945	-0.591
Last Notable Activity_Olark Chat Conversation	-0.7672	0.329	-2.329	0.020	-1.413	-0.121
Last Notable Activity_Unreachable	1.8374	0.523	3.511	0.000	0.812	2.863

Model Results

Sensitivity/True positive rate/Recall: 68%

Specificity: 89.5%

False Positive Rate: 10.5%

Positive predictive value/Precision: 80%

Negative Predicted Value: 82%

If we choose cutoff as 0.52, Final Precision of model is 80%

Results

- We achieved accuracy of around 80% on both Train & Test data.
- The leads that we predicted converted i.e. Hot Leads have achieved Precision of 73% and 74% on our Train & Test data.
- The trade-off between True Positive Rate and False Positive Rate is also good. The area under the ROC curve is around 0.89

Answer to Subjective Questions

Q1: Top three variables in your model which contribute most towards the probability of a lead getting converted:

- Total Time Spent on Website
- Lead origin
- What is your current occupation

Q2: Top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion:

- Lead Origin_Lead Add Form
- Last Activity_Had a Phone Converstion
- What is your current occupation_Working Professional

Answer to Subjective Questions

Q3: 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. Suggest a good strategy they should employ at this stage.

Ans: Here with team of 10 interns we can afford to make higher number of calls to potential leads so we will change the cutoff to lower value so that false negative rate is minimized, and we can utilize the same to get higher conversion. Also we can offer some scholarship, 0% EMI option if the prospective lead is financially not strong.

Answer to Subjective Questions

Q4: 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. Suggest a strategy they should employ at this stage.

Ans: Here the team wants to minimize the no. of useless phone calls so we will change the cutoff to higher value so that false positive rate is minimized and team can call only the hot leads and utilize their time in other things.

Conclusions

- Some of the driving variables to increase the hot-leads are “What is your current occupation” and “Lead origin”.
- With the current model, the sales team can now more focus on potential leads instead of earlier making useless phone calls and generate more business.
- We can have more hot leads now as our model precision is 80%.
- Education company can provide scholarships and other offers (like 0% EMI scheme) to get more potential leads.