LEAD SCORING CASE STUDY

BUILDING LOGISTIC REGRESSION MODEL TO PREDICT WHETHER A LEAD FOR ONLINE COURSES FOR AN EDUCATION COMPANY WOULD BE SUCCESSFULLY CONVERTED OR NOT

Nishal s devadiga
Tushar

PROBLEM SOLVING METHODOLOGY

- Understanding the data
- Applying RFE to identify the best performing features
- Building the model with features selected with RFE. Eliminate features with high VIF and p-value
- Prerform model evaluation metrics like sensitivity, specificity, precision, recall, etc.
- Decide the probability threshold based on optimal cutoff
- Use model for predictions on test dataset and perform model evaluation for the test set.

BUSINESS OBJECTIVES

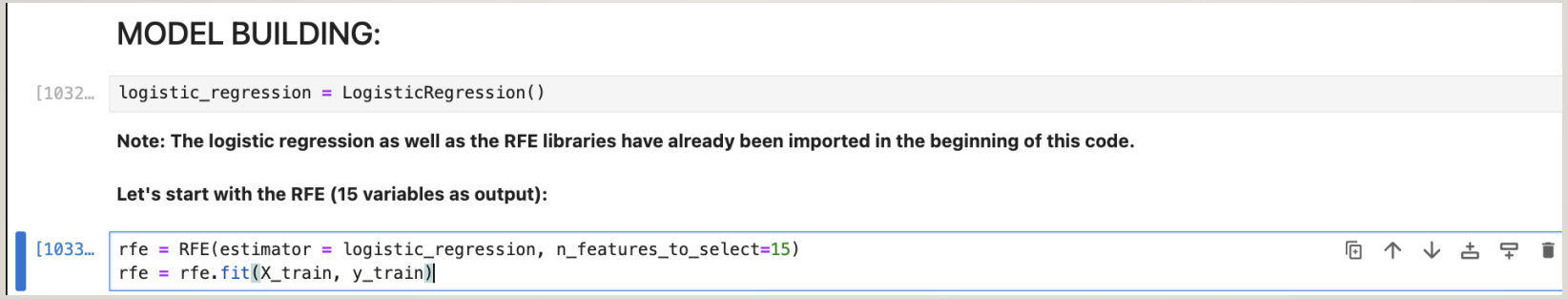
- Help Education company to select the most promising leads.
- To build a logistic regression model to assign a lead score value between 0 and 100 to each of the leads which can be used by the company to target potential leads
- Building logistic regression model to predict the lead conversion probabilities for each lead.
- Decide on a probability threshold value above which a lead will be predicted as converted, whereas not converted if it is below it.

DATA PREPARATION

- There are some columns which have select values. This may be optional values , these are converted as nulls.
- Removing all the columns that are not required and have 35% null values
- Dummies are created for categorical variables.
- Specialization, What matters most to you in choosing a course, Country, What is your current occupation in these null values are replaced as ‘not provided’. These features are important for our analysis and we cannot drop them.

FEATURE SELECTION USING RFE

- Recursive feature elimination is an optimization technique for finding the best performing subset of features.



BUILDING THE MODEL

Building the final model using the features selected by RFE.
13 features are selected after elimination of variables having High VIF and high p-value.

	Features	VIF
1	Total Time Spent on Website	2.33
0	TotalVisits	2.28
4	Lead Source_google	2.04
3	Lead Source_direct traffic	1.91
5	Lead Source_organic search	1.60
9	Last Activity_sms sent	1.49
2	Lead Origin_lead add form	1.47
6	Lead Source_welingak website	1.31
10	What is your current occupation_working profes...	1.17
7	Do Not Email_yes	1.10
8	Last Activity_olark chat conversation	1.02
12	Last Notable Activity_unreachable	1.01
11	Last Notable Activity_had a phone conversation	1.00

PREDICTING THE CONVERSION PROBABILITY

- Creating new column predicted with 1 if coversion_prob is > 0.5 else 0.

[1048...

	Converted	Conversion_Prob	Predicted
0	1	0.610084	1
1	0	0.222737	0
2	0	0.424993	0
3	0	0.222737	0
4	0	0.433499	0

Since the conversion is now done with substitutions of 0 and 1 (w

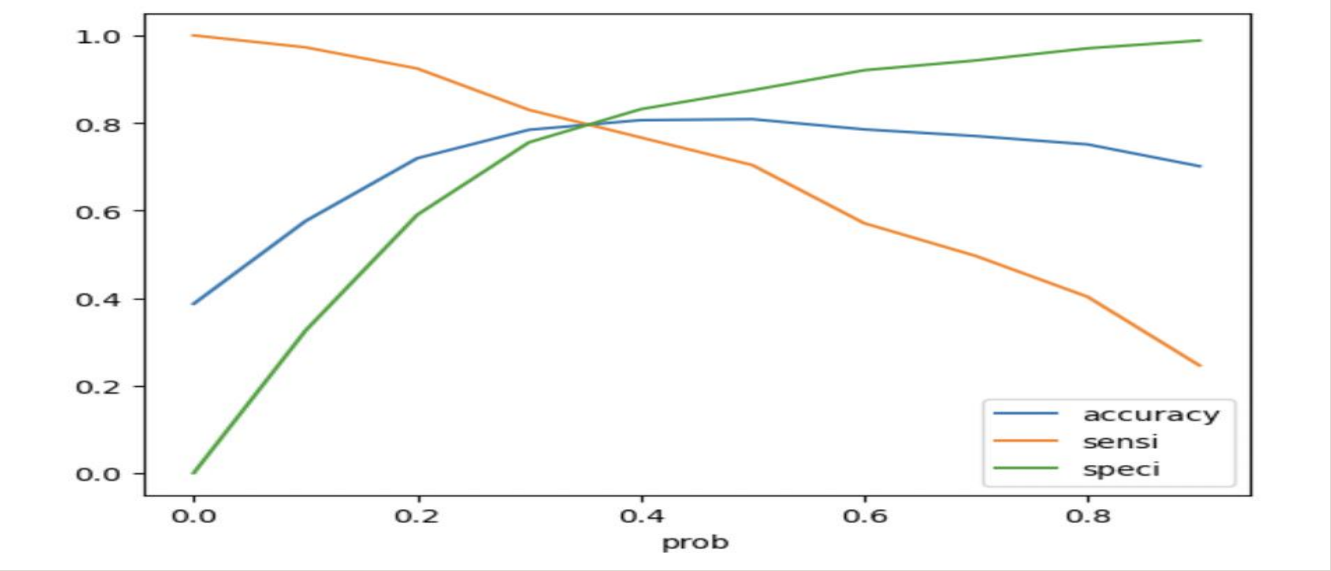
LEAD SCORE

- Lead_score is calculated by coversion_prob*100

	Converted	Conversion_Prob	final_predicted	lead_score
0	0	0.342237	0	34.223696
1	1	0.849376	1	84.937558
2	1	0.982499	1	98.249944
3	1	0.823578	1	82.357820
4	0	0.071354	0	7.135377

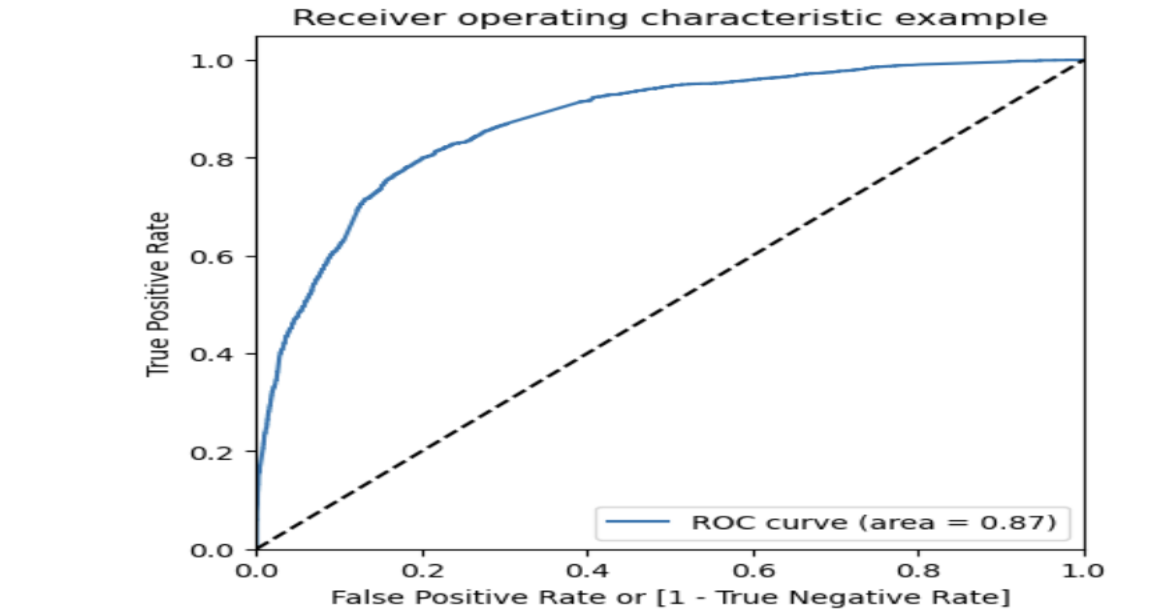
FINDING OPTIMAL PROBABILITY THRESHOLD

- Accuracy, sensitivity and specificity is calculated for various values of probability threshold and plotted in the graph.
- From the curve, 0.38 is found to be the optimum point of cutoff probability.



PLOTTING THE ROC CURVE

- It shows the tradeoff between sensitivity and specificity(increase in sensitivity is accompanied by a decrease in specificity).



The graph above indicates that the area under the ROC curve is 0.88. T

RECOMMENDATION AND PROBLEM SOLUTION

- 1. Which are minimize three variables in your model which contribute most towards the probability of a lead getting converted?
- Answer: As per the coefficients of the model, 1) Last Notable Activity_had a phone conversation. 2) TotalVisits. 3) Total Time Spent on Website.
- 2. What are the top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion?
- Answer: 1) Last Notable Activity_had a phone conversation. 2)What is your current occupation_working professional. 3)Lead Origin_lead add form.

RECOMMENDATION AND PROBLEM SOLUTION

- 3. 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.
- Answer: For 2 months, we need to lower the probability threshold as sensitivity is high for lower probability, to increase the number of lead classified as 1. Probability threshold can be 0.4 or even 0.3.
- 4. 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.
- Answer: We can increase the probability threshold and as a result specificity increases. Which can minimize the rate of useless phone calls and sales team can focus on new work as well