upGrad

Lead Scoring Case Study

-Sourabh and Nikhil

Problem statement

- To build a Logistic Regression Model to predict whether a lead for online courses for an education company named X Education would be successfully converted or not.

Business Objective:

- To help X Education to select the most promising leads(Hot Leads), i.e. the leads that are most likely to convert into paying customers.
- 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.

The objective is thus classified into the following sub-goals:

- Create a 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.

- Multiply the Lead Conversion probability to arrive at the Lead Score value for each lead.

Solution methodology for arriving at the final goal

- A complete study and analysis based on the given Leads dataset.
- Applying Recursive feature elimination to identify the best performing subset of features for building the model.
- Building the model with features selected by RFE. Eliminate all features with high p-values and VIF values and finalize the model
- Perform model evaluation with various metrics like sensitivity, specificity, precision, recall, etc.
- Decide on the probability threshold value based on Optimal cutoff point and predict the dependent variable for the training data.
- Use the model for training on the train set and then prediction on the test dataset and perform model evaluation on both.

Data Preparation and Feature Engineering

upGrad

Remove columns which has only one unique value

Deleting the following columns as they have only one unique value and hence cannot be responsible in predicting a successful lead case – 'Magazine', 'Receive More Updates About Our Courses' , 'Update me on Supply Chain Content' , 'Update me on Supply Chain Content' and 'I agree to pay the amount through cheque'.

Removing rows where a particular column has low missing values

'Lead Source' is an important column for analysis. Hence all the rows that have null values for it were dropped.

Imputing NULL values with Median

The columns 'TotalVisits' and 'Page Views Per Visit' are continuous variables with outliers. Hence the null values for these columns were imputed with the column median values.

Contd...

Imputing NULL values with Mode

The columns ‘Country’ is a categorical variable with some null values. Also majority of the records belong to Country ‘India’. Thus imputed the null values for this with mode(most occuring value). Then binned rest of category into 'Outside India‘.

Handling 'Select’ values in some columns

There are some columns in dataset which have a value called 'Select’. The Select values in columns were converted to Nulls.

Outlier Treatment

Outliers were removed using IQR method

Recursive Feature Elimination process is applied until all the features in the dataset are exhausted. Features are then ranked according to when they were eliminated.

Running RFE with the output number of the variable equal to 15

Dummy variables were created for all categorical variables and original variables were dropped.

Splitting of data into Train-Test data sets.

The original dataframe was split into train and test dataset. The train dataset was used to train the model and test dataset was used to evaluate the model.

Feature scaling of continuous variables using Standard Scalar method was performed.

Building the Model

Generalized Linear Models from StatsModels is used to build the Logistic Regression model.

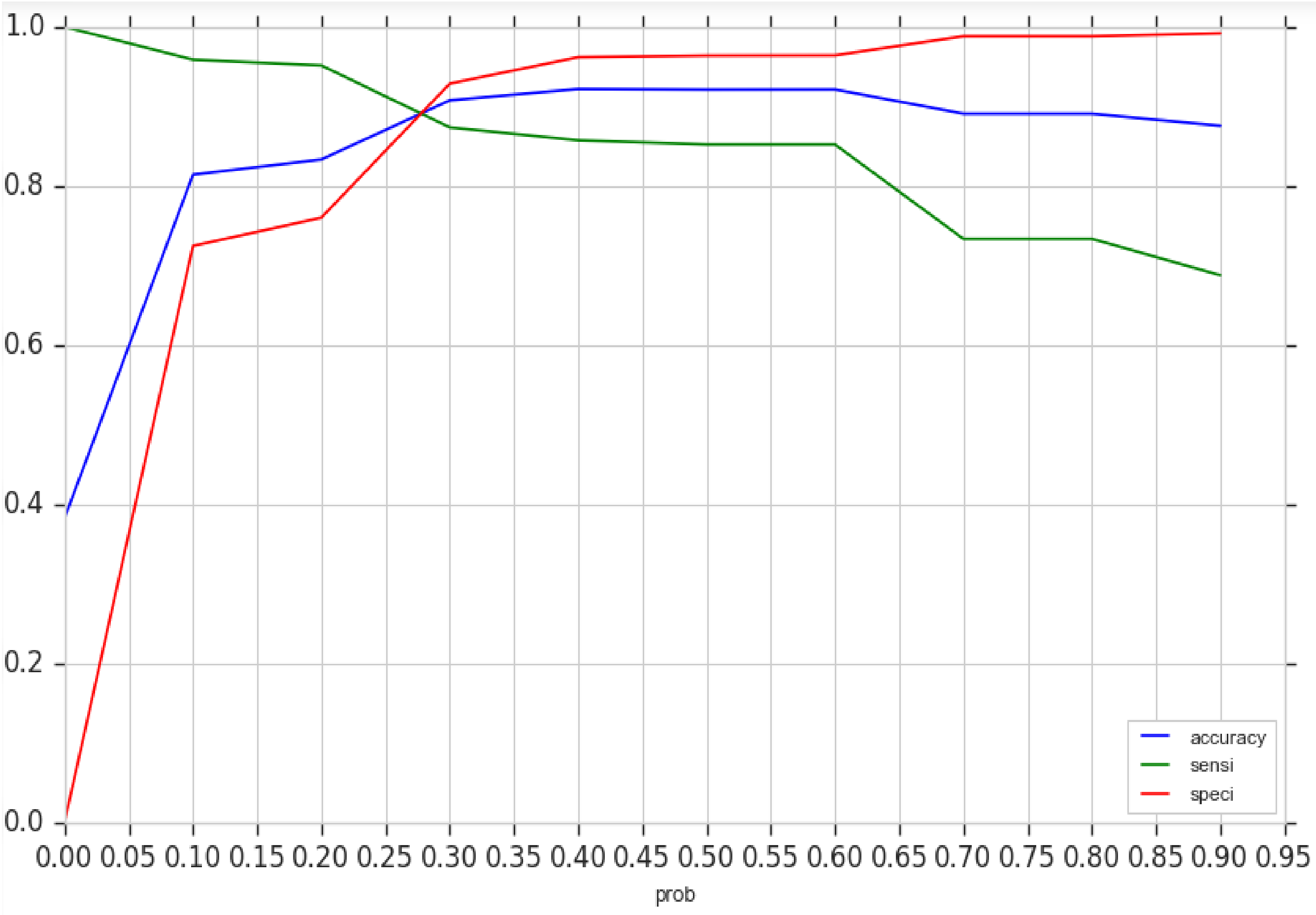
- The model is built initially with the 15 variables selected by RFE.
- Unwanted features are dropped serially after checking p values (<0.5) and VIF (< 5) and model is built multiple times.
- The final model with 13 features, passes both the significance test and the multi-collinearity test.

Finding Optimal Probablility Threshold

Optimal cutoff probability is that prob where we get balanced sensitivity and specificity.
The accuracy sensitivity and specificity was calculated for various values of probability threshold and plotted in the graph below:

From the graph, 0.28 is found to be the optimum point for cutoff probability.

At this threshold value, all the 3 metrics – accuracy sensitivity and specificity was found to be well above 80% which is a well acceptable value.



ROC and AUC

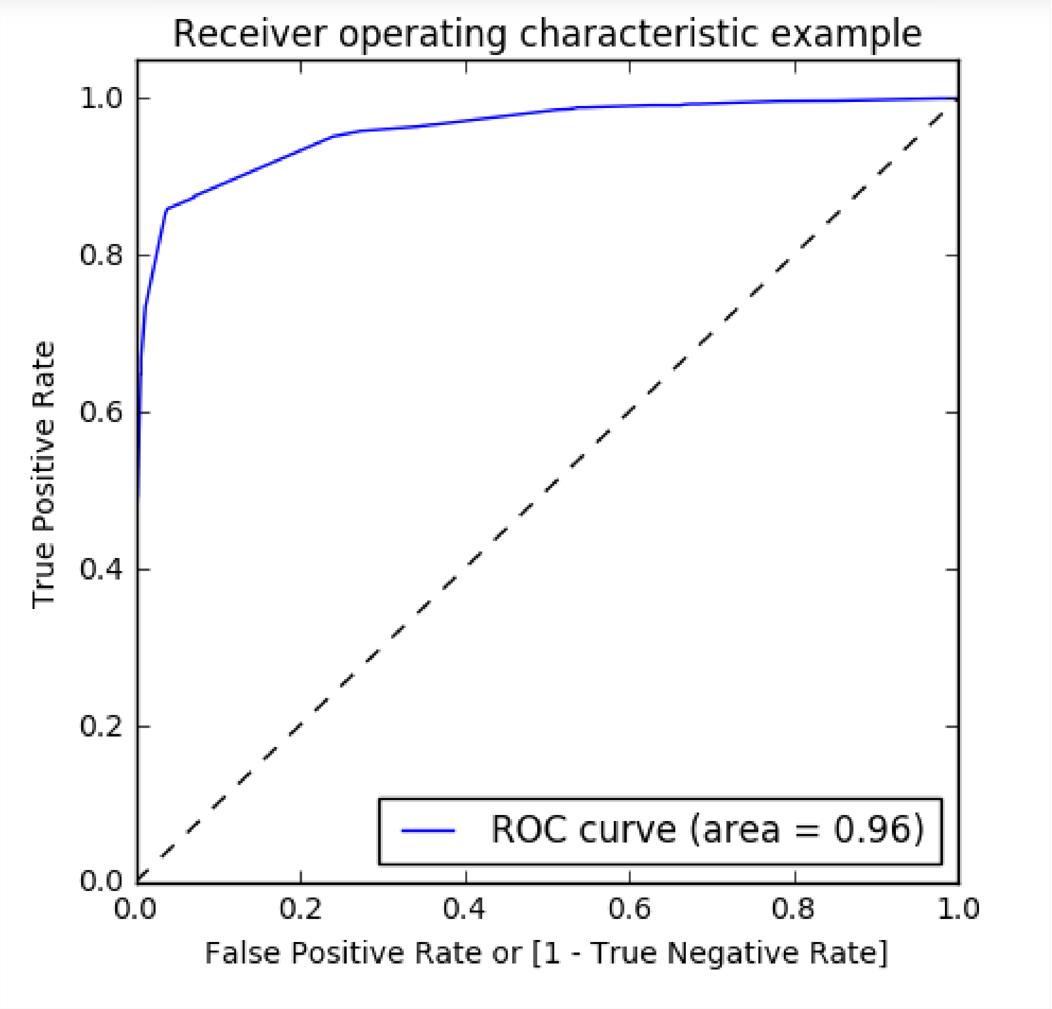
Receiver Operating Characteristics (ROC) Curve

It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).

Area under the Curve (GINI)

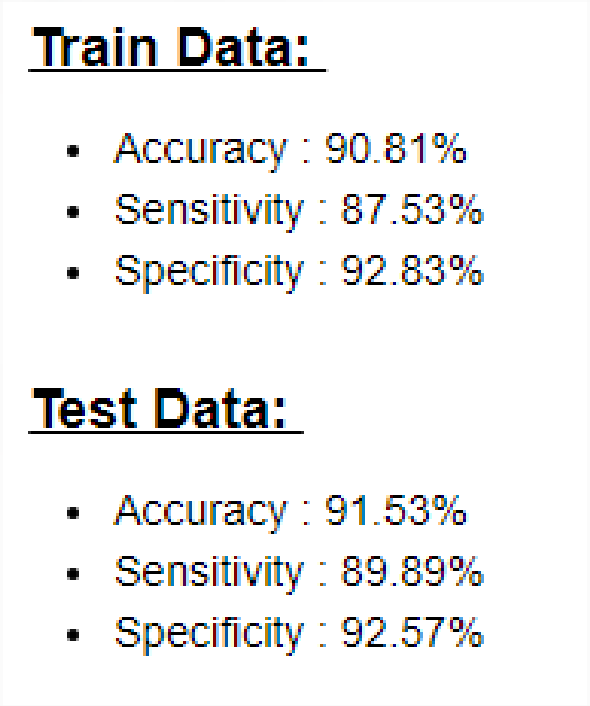
By determining the Area under the curve (AUC) of the ROC curve, the goodness of the model is determined. Since the ROC curve is more towards the upper-left corner of the graph, it means that the model is very good. The larger the AUC, the better is the model.

- The value of AUC for our model is 0.96



Accuracy, Sensitivity and Specificity

Following figures were obtained after running the model on the Train and Test dataset:



Train Data:

- Accuracy : 90.81%
- Sensitivity : 87.53%
- Specificity : 92.83%

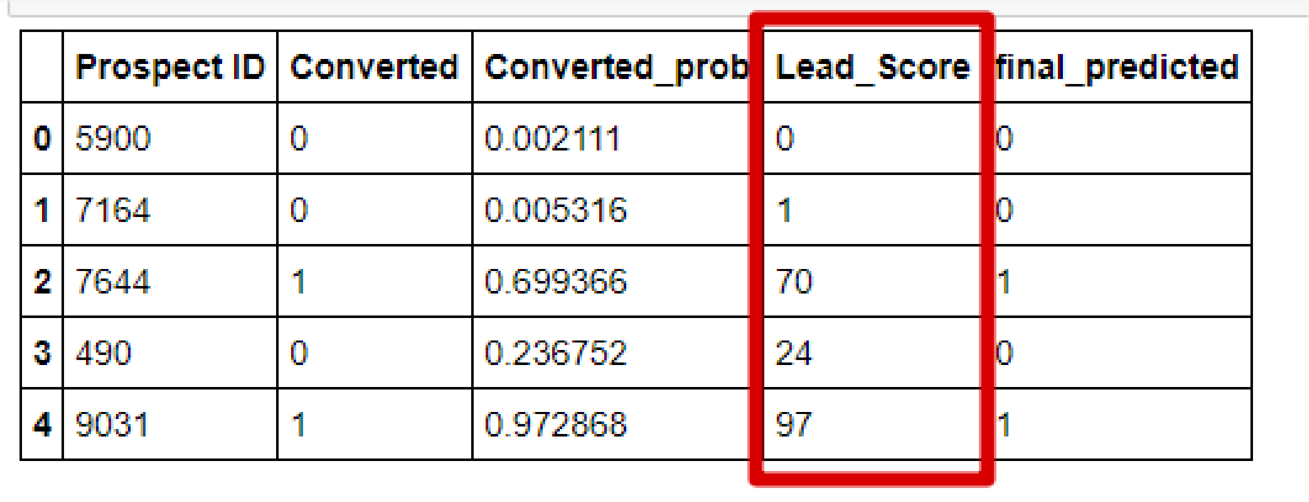
Test Data:

- Accuracy : 91.53%
- Sensitivity : 89.89%
- Specificity : 92.57%

Lead Score calculation

Lead Score is calculated for all the leads in the original dataframe.

Formula for Lead Score calculation is:
Lead Score = 100 * Conversion Probability



	Prospect ID	Converted	Converted_prob	Lead_Score	final_predicted
0	5900	0	0.002111	0	0
1	7164	0	0.005316	1	0
2	7644	1	0.699366	70	1
3	490	0	0.236752	24	0
4	9031	1	0.972868	97	1

The Conversion Probability is multiplied by 100 to obtain the Lead Score for each lead.

Higher the lead score, higher is the probability of a lead getting converted and vice versa.

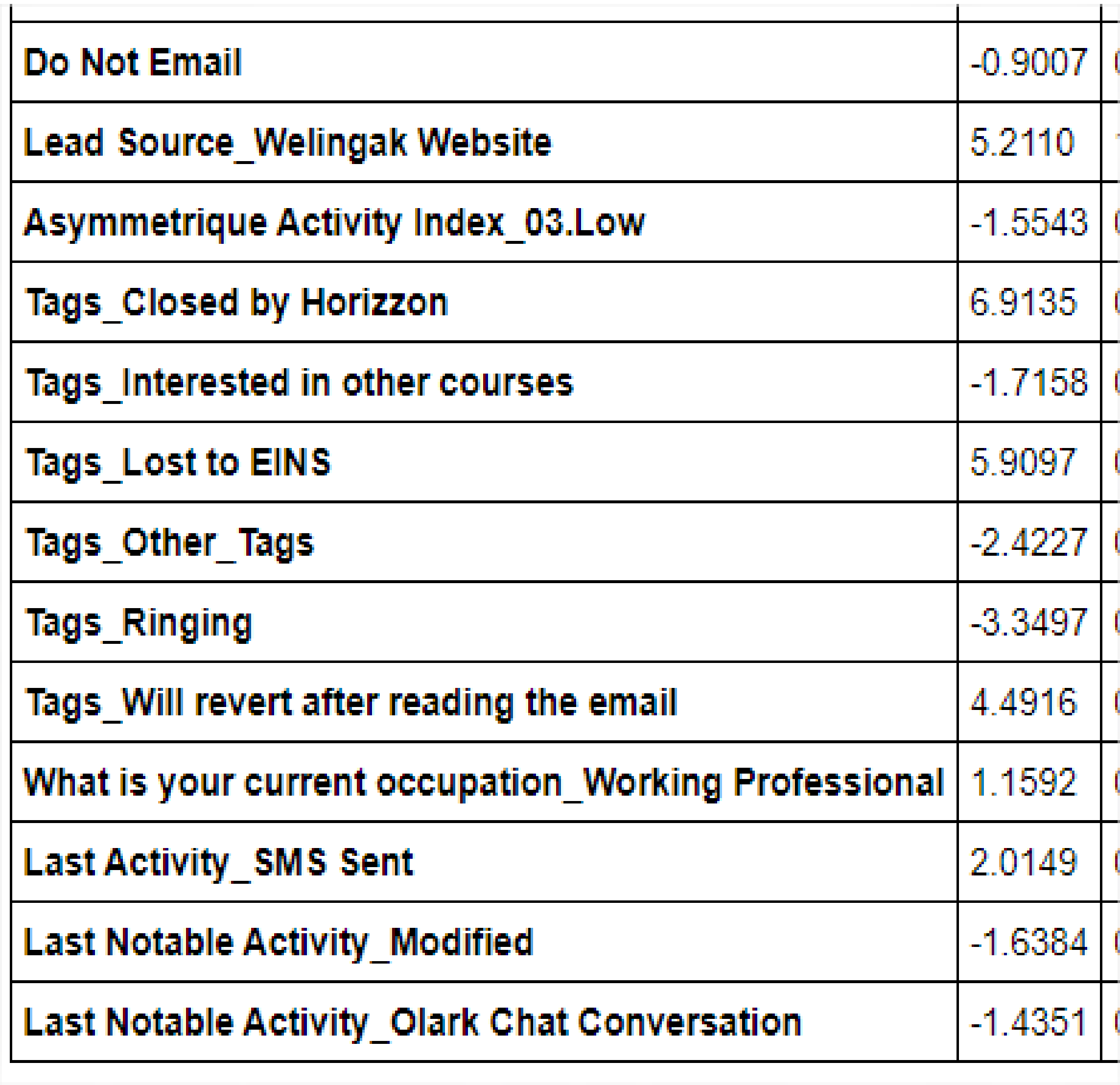
Determining Feature Importance

13 features have been used by our model to successfully predict if a lead will get converted or not.

- The Coefficient (beta) values for each of these features from the model parameters are used to determine the order of importance of these features.

- Features with high positive beta values are the ones that contribute most towards the probability of a lead getting converted.

- Similarly, features with high negative beta values contribute the least.



Do Not Email	-0.9007
Lead Source_Welingak Website	5.2110
Asymmetrique Activity Index_03.Low	-1.5543
Tags_Closed by Horizzon	6.9135
Tags_Interested in other courses	-1.7158
Tags_Lost to EINS	5.9097
Tags_Other_Tags	-2.4227
Tags_Ringing	-3.3497
Tags_Will revert after reading the email	4.4916
What is your current occupation_Working Professional	1.1592
Last Activity_SMS Sent	2.0149
Last Notable Activity_Modified	-1.6384
Last Notable Activity_Olark Chat Conversation	-1.4351

Inferences

After trying several models, we finally chose a model with the following characteristics:

- All variables have p-value < 0.05

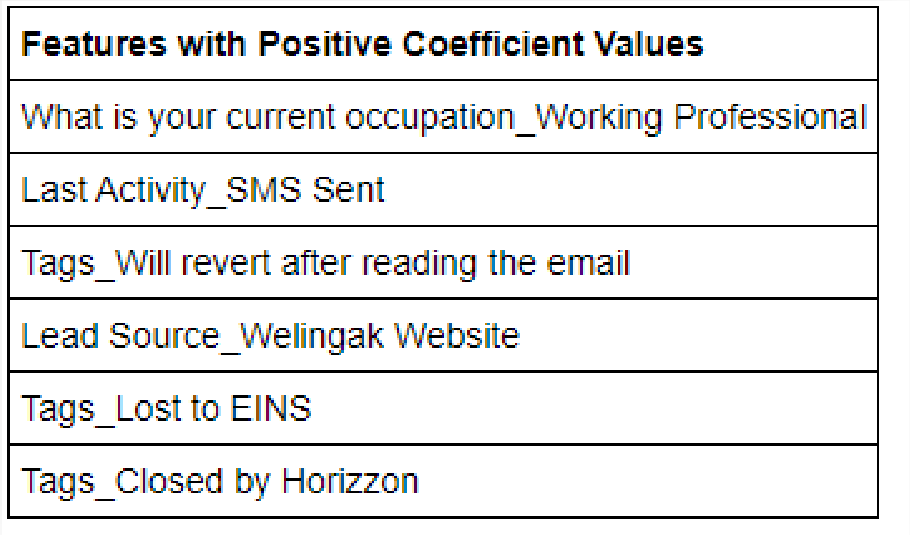
- All the features have very low VIF values, meaning, there is hardly any multicollinearity among the features

- The overall accuracy of 0.9056 at a probability threshold of 0.28 on test dataset is acceptable.

Inferences

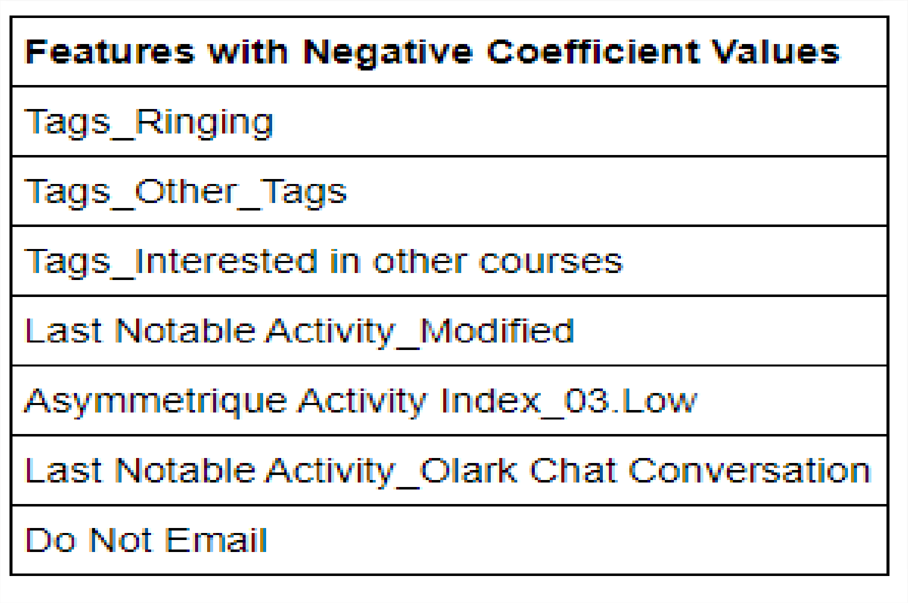
Based on our model, some features are identified which contribute most to a Lead getting converted successfully.

The conversion probability of a lead increases with increase in values of the following features in descending order →



Features with Positive Coefficient Values
What is your current occupation_Working Professional
Last Activity_SMS Sent
Tags_Will revert after reading the email
Lead Source_Welingak Website
Tags_Lost to EINS
Tags_Closed by Horizzon

The conversion probability of a lead increases with decrease in values of the following features in descending order →



Features with Negative Coefficient Values
Tags_Ringing
Tags_Other_Tags
Tags_Interested in other courses
Last Notable Activity_Modified
Asymmetrique Activity Index_03.Low
Last Notable Activity_Olark Chat Conversation
Do Not Email

Recommendations and Problem solutions

upGrad

The top variables in our model which contribute most towards the probability of a lead getting converted are:

- ‘Tag’
- ‘Lead Source’
- ‘What is your current occupation’
- ‘Last Activity’

are the top variables that contribute most towards the probability of a lead getting converted.

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 are:

- Tags_Closed by Horizzon
- Tags_Lost to EINS
- Lead Source_Welingak Website

Recommendations and Problem solutions Contd…

upGrad

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.

➡ We will choose a lower threshold value for Conversion Probability. This will ensure the Sensitivity rating is very high which in turn will make sure almost all leads that are likely to Convert are identified correctly and the agents can make phone calls to as much of such people as possible.

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.

➡ We will choose a higher threshold value for Conversion Probability. This will ensure the Specificity rating is very high, which in turn will make sure almost all leads that are on the brink of the probability of getting Converted or not are not selected. As a result the agents won't have to make unnecessary phone calls and can focus on some new work.

Thank you