Lead scoring case study

Business Objective

. BUSINESS OBJECTIVE:

- To help the Online Education company X to identify the most promising 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.

Thus the above objective is achieved by attaining the following sub-goals:

a) Logistic Regression is used to classify the lead and predict the lead conversion probabilities

b) To decide the probability threshold value for lead conversion

c) Finally to derive the lead score multiply the lead conversion probability with 100

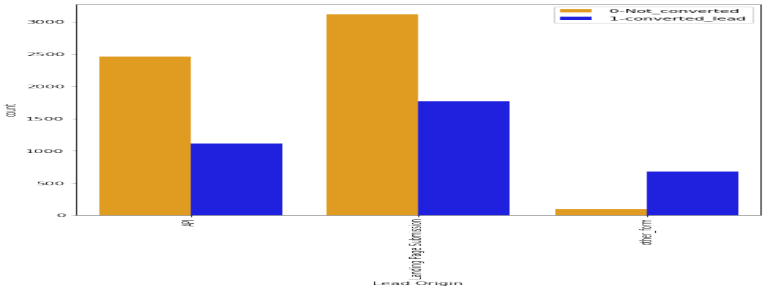
Data Preparation

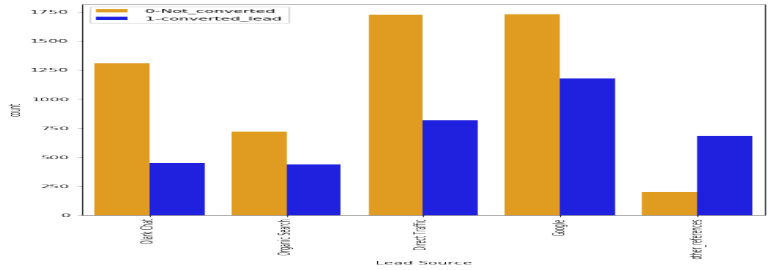
Identify the columns with “Select” as input and replace them with NAN values	Some categorical columns('Lead Profile, City etc) have values as "Select". This is basically a by-default field within the application, whenever the user doesn't populate these fields while filling our forms.
Remove the columns that have skewed distribution	Columns such as 'Do Not Email','Do Not Call','Country','What matters most to you in choosing a course','I agree to pay the amount through cheque','A free copy of Mastering The Interview’ etc.
Removing the columns where missing values are greater than 45%	Columns having high missing percentage such as 'Asymmetrique Activity Index','Asymmetrique Profile Index', 'Asymmetrique Activity Score','Asymmetrique Profile Score’ 'Lead Quality','Lead Profile','How did you hear about X Education'
Imputing Null values with appropriate measure such as median and mode	The columns such as ‘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.

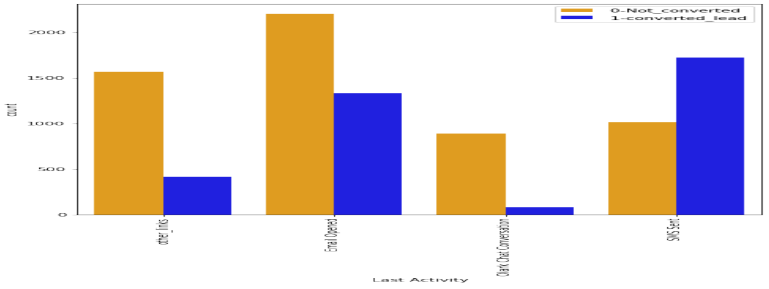
Data Preparation

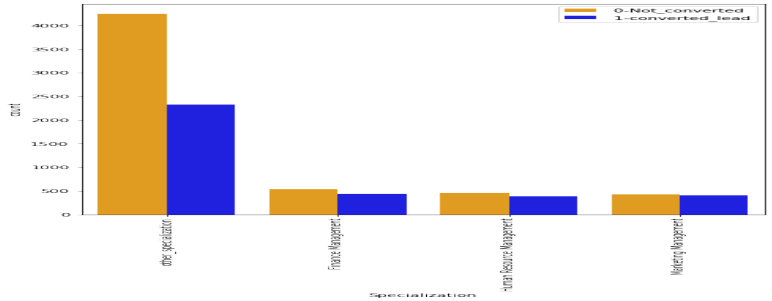
For some columns we have created separate missing category to avoid skewness	The columns such as What is your occupation represents the occupation of the lead However, imputing this the mode might create inaccuracies while building predicting model. Hence, we can impute the missing values with 'missing'.
Assigning a Unique Category to NULL/SELECT values	All the missing values in the columns were binned into a separate column 'others'. Instead of deleting such columns, this strategy adds more information into the dataset and results in the change of variance
Handling Outliers	The outliers were present in the columns 'TotalVisits' & 'Page Views Per Visit' . The removing of outliers could result in the loss og information about the potential lead. Therefore, we have capped the outliers.
Creating dummies	Creating dummies for categorical variables

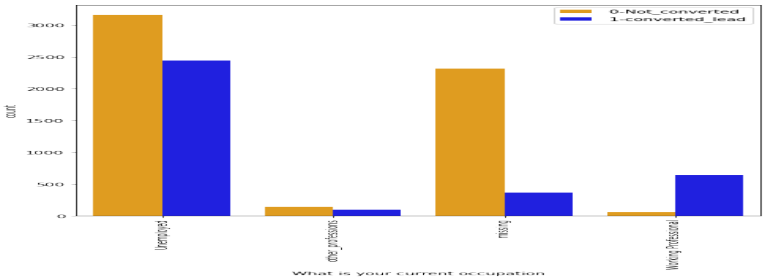
Data Visualization

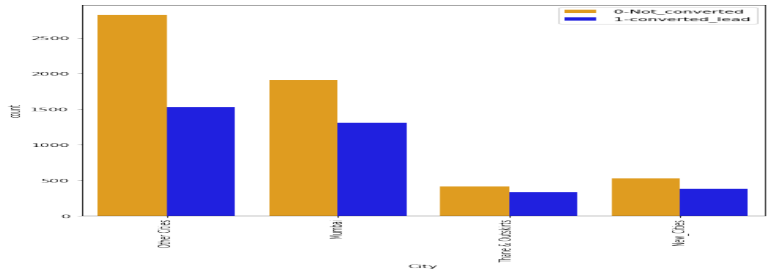


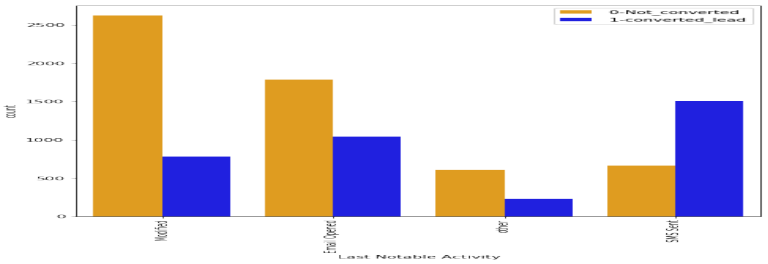












1.Lead Origin : API and Landing Page Submission have more conversion rate as compared to others. However, Lead Add Form has max conversion rate but count of lead are not very high.
2.Lead Source: Google and Direct traffic generates maximum number of leads.
3.Last_activity and Last Notable activity : The conversion of sms sent is very high.
4.What is your current occupation : The conversion rate of working professionals are high

Model Building

Test-Train Split and Feature Scaling

- The 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.
- 'Standardization' was used to scale the data for modelling. It basically brings all of the data into a standard normal distribution with mean at zero and standard deviation one.

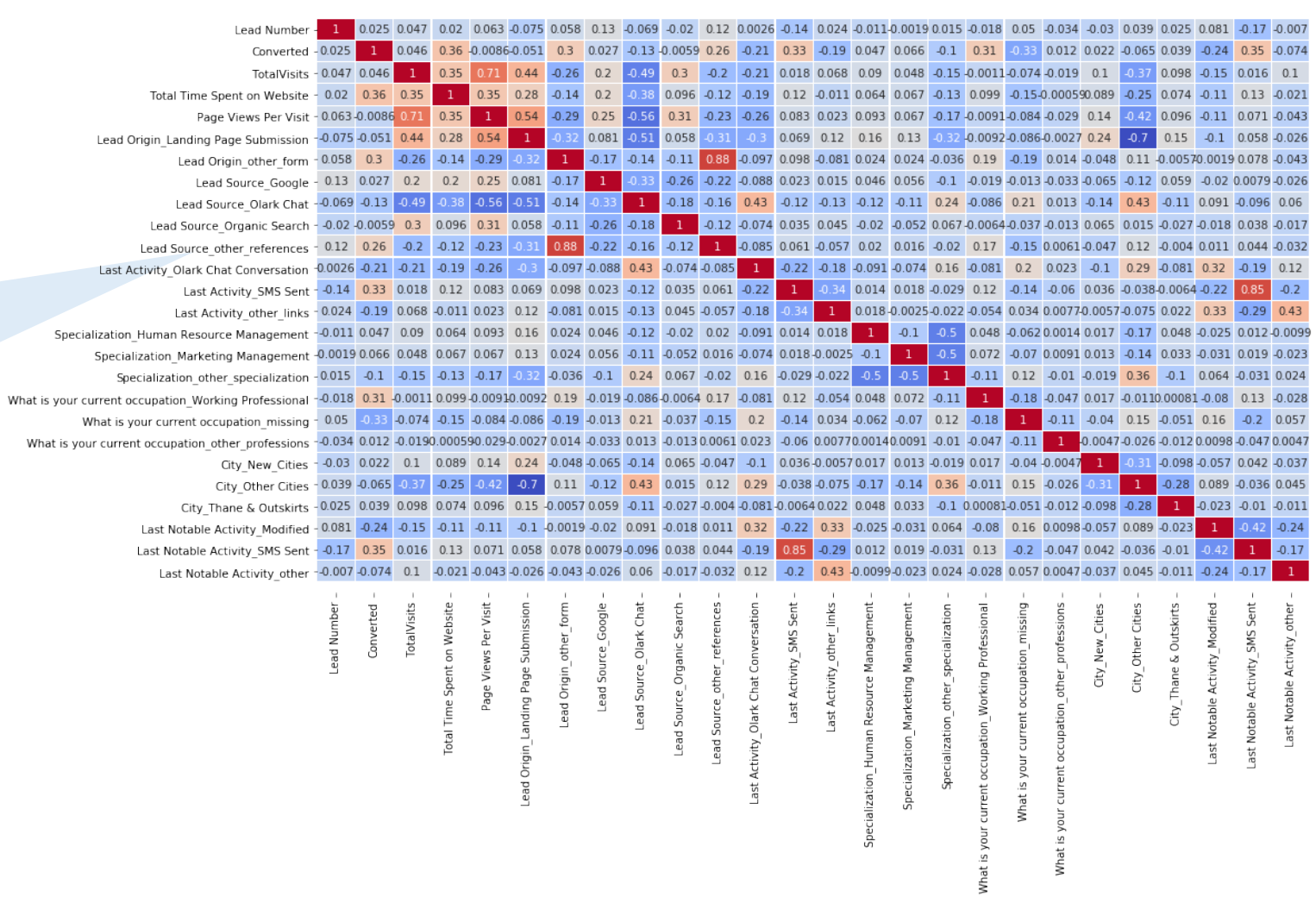
Feature Selection using RFE

- Recursive feature elimination is an optimization technique for finding the best performing subset of features.
- Running RFE with the output number of the variable equal to 20.

Model Building

Looking at the Correlation Matric through Heatmap

It clearly indicates the highly correlated variables i.e Lead Origin_other_form','Last Activity_SMS Sent' that need to be drop



Final Model

- GLM from StatsModels is used to build the Logistic Regression model
- The model is built initially with the 20 variables selected by RFE.
- Unwanted features are dropped serially after checking p values (< 0.5) and VIF (< 3) and model is built multiple times
- The final model with 15 features, passes both the significance test and the multi-collinearity test.

FEATURES	coef	P>|z|
const	-0.5621	0
TotalVisits	0.2137	0
Total Time Spent on Website	1.0333	0
Page Views Per Visit	-0.3791	0
Lead Origin_Landing Page Submission	-0.5682	0
Lead Source_Google	0.4237	0
Lead Source_Olark Chat	0.9972	0
Lead Source_Organic Search	0.2668	0.038
Lead Source_other_references	2.1954	0
Last Activity_Olark Chat Conversation	-1.7263	0
Last Activity_other_links	-1.2195	0
What is your current occupation_Working Professional	2.609	0
What is your current occupation_missing	-1.1752	0
City_Other Cities	-0.4554	0
Last Notable Activity_SMS Sent	1.3099	0
Last Notable Activity_other	0.8297	0

FEATURES	VIF
City_Other Cities	2.81
Page Views Per Visit	2.63
Lead Source_Olark Chat	2.61
TotalVisits	2.23
Lead Origin_Landing Page Submission	2.17
Last Activity_other_links	1.85
Lead Source_Google	1.76
What is your current occupation_missing	1.61
Last Notable Activity_SMS Sent	1.55
Last Activity_Olark Chat Conversation	1.54
Lead Source_Organic Search	1.50
Last Notable Activity_other	1.49
Lead Source_other_references	1.38
Total Time Spent on Website	1.27
What is your current occupation_Working Profes...	1.17

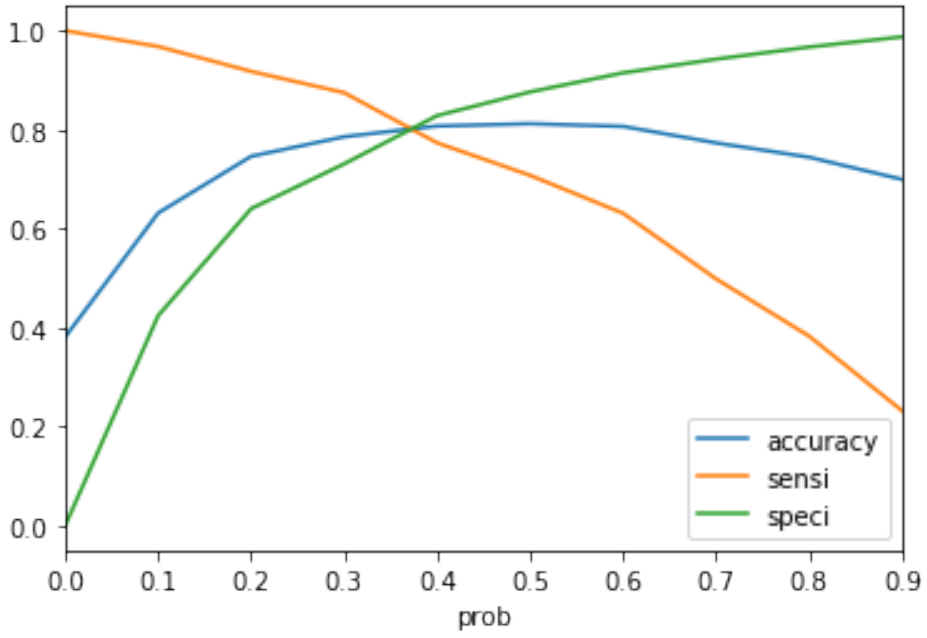
Predicting the Conversion Probability and Predicted Columns

- Creating new column 'predicted' with 1 if Conversion_Prob > 0.5 else 0
- Showing top 5 records of the dataframe in the below table

Index	Conversion	Conversion_prob	Lead Number	Predicted
0	0	0.329506	1871	0
1	0	0.219320	6795	0
2	0	0.362821	3516	0
3	0	0.824296	8105	1
4	0	0.329506	3934	0

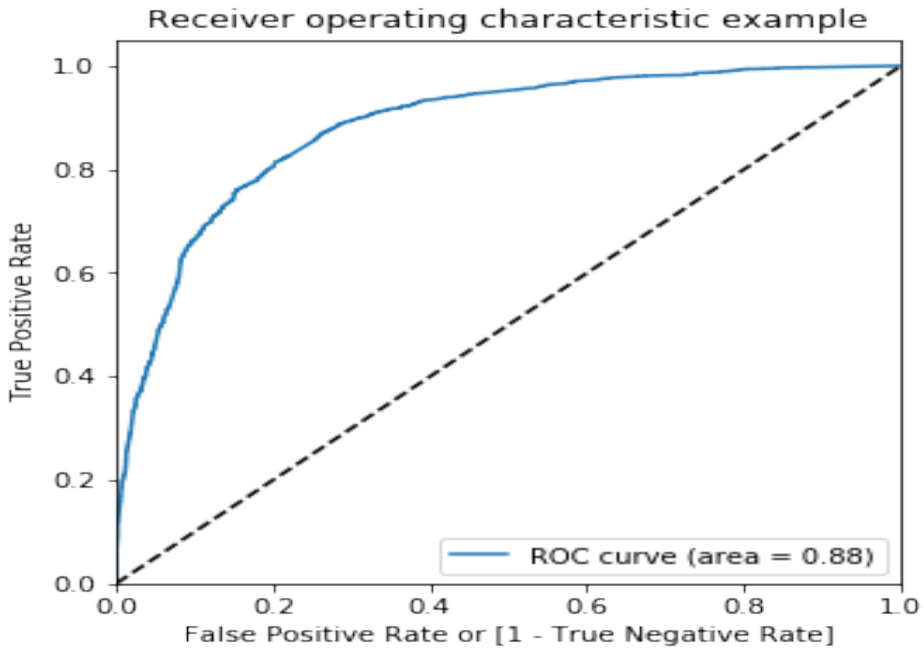
Optimal Probability Threshold and ROC

In the below plot, we can observe that specificity (green line) starts at 0 and goes very high. While senstivity follows reverse trend it is 1 at the lower end.Beyond the value of 0.2, the accuracy value is not much changing and anything above this should be convenient.



Optimal point =0.4

ROC curve as above shows the trade of between senstivity and specificity. X axis has true positive rate; Y axis as False positive rate. he below ROC curve has an area of 0.88 which can be considered as decent.



ROC

Evaluating the model on the trained dataset

CONFUSION MATRIX

# Predicted #Actual	Not converted	Converted
Not Converted	3507	495
Converted	721	1745

PROBABILITY THRESHOLD = 0.4

OTHER MEASURES (Based on test data)

Accuracy = 0.80

Sensitivity=0.77

Specificity= 0.82

Precision = 0.77

Recall= 0.70

Making Prediction on the Test Data

- The final model on the train dataset is used to make predictions for the test dataset
- The train data set was scaled using the scaler.transform function that was used to scale the train dataset.
- The Predicted probabilities were added to the leads in the test dataframe.
- Using the probability threshold value of 0.4, the leads from the test dataset were predicted if they will convert or not.
- Showing top 5 records of the dataframe in the below table

Index	Conversion	Conversion_prob	Lead Number	Predicted
0	1	0.4149	4269	1
1	1	0.8578	2376	1
2	1	0.6919	7766	1
3	0	0.0804	9199	0
4	1	0.7197	4359	1

Evaluating the model on the test dataset

CONFUSION MATRIX

# Predicted #Actual	Not converted	Converted
Not Converted	1384	293
Converted	250	845

PROBABILITY THRESHOLD = 0.4

OTHER MEASURES

- Accuracy = 0.80
- Sensitivity=0.77
- Specificity= 0.82

Lead Score Calculation

*LEAD SCORE = 100 * CONVERSION PROBABILITY*

Index	Converted	Converted_Prob	Lead Number	predicted	Lead_Score
0	0	0.329506	1871	0	33
1	0	0.219320	6795	0	22
2	0	0.362821	3516	0	36
3	0	0.824296	8105	1	82
4	0	0.329506	3934	0	33

Final Analysis : Determining the Feature Importance

- In our model, 15 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.

1. What is youre current occupation_Working Professional
2. Lead_Source_Other_Referecnes
3. Last Notable Activity_other

FEATURES	Coefficient
const	-0.5621
TotalVisits	0.2137
Total Time Spent on Website	1.0333
Page Views Per Visit	-0.3791
Lead Origin_Landing Page Submission	-0.5682
Lead Source_Google	0.4237
Lead Source_Olark Chat	0.9972
Lead Source_Organic Search	0.2668
Lead Source_other_references	2.1954
Last Activity_Olark Chat Conversation	-1.7263
Last Activity_other_links	-1.2195
What is your current occupation_Working Professional	2.6090
What is your current occupation_missing	-1.1752
City_Other Cities	-0.4554
Last Notable Activity_SMS Sent	1.3099
Last Notable Activity_other	0.8297

THANK YOU