Lead Scoring Case Study

Sharod Dey

Agenda

Problem Statement and Overall Approach

Data Preparation and EDA

Model Building

Model Evaluation

Interpretation and Recommendations




Introduction

In this presentation we will go through the problem and the approach taken to develop the solution. We will also explore some salient points while building the models. Finally, we will present actionable insights and recommendations for the business.

Problem Statement and Overall Approach

Analysis of Leads for an education company

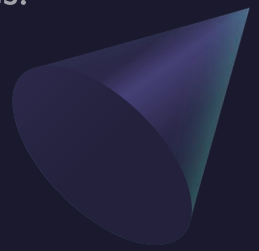
- An education company, X Education, markets its products across various platforms. When someone interacts with these marketing contents they are classified as a lead.

- The issue is to identify which of these leads are most likely to convert into a customer who will actually consume the product, which is this case are courses. This identification of potential leads is important to optimally spend resources in pursuing the leads which are most likely to be converted. Therefore, reducing costs and maximising revenue.

- Present conversation rate of leads is at 30%. The objective of this solution is to increase the conversion rate as per the target received of 80%. This is to be achieved by building a model which will assign a score to each lead, higher scores will indicate higher likelihood of conversion.

- The approach to this problem would be to build a Logistic Regression model and assign the probabilities obtained as the lead score. This model will also enable us to interpret the driver variables easily and be able to provide the business with practical recommendations. We will also an additional binary recommendation of Yes/No along with the score for ease of comprehension.

- Since we are focusing on identifying which leads are more likely to convert, the metric we shall focus on to evaluate our model will be Recall or Sensitivity. A higher recall will allow the model to identify most leads which are likely to convert, meaning reduce the false negatives and increase the true positives.

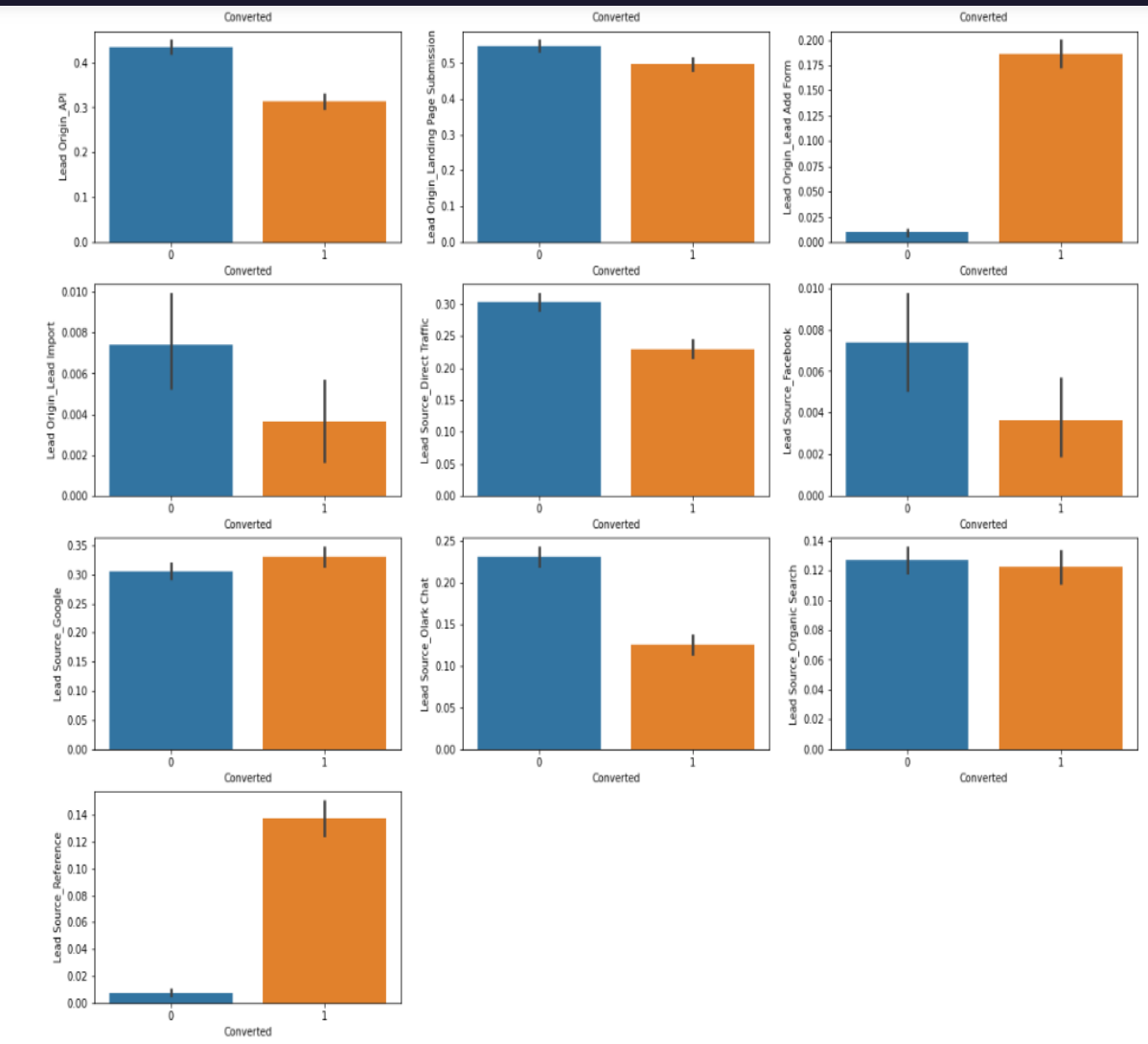


Data Preparation and EDA

Null value and outliers handling

Univariate and Bivariate analysis

- The first step was to handle the null values. This was done by calculating the percentage of null value in each column of the dataset.
- Columns with high null value percentages (>20%) were dropped. Following this, the remaining variables had low percentage of null values and this was handled by imputation. Mode was used to impute the categorical variables and mean was used for continuous.
- Next step was to check for outliers in the continuous numerical variables. This was done by checking the 95^{th}, 99^{th} and 100^{th} percentiles and plotting boxplots. The identified outliers were dropped.
- The binary categorical variables (Yes/No) were encoded to 1/0. Dummy variables were created for the remaining categorical variables after some value grouping and further cleaning. Dummy variables were One Hot encoded and redundant columns were dropped.
- Visualizations were created for univariate and bivariate analysis. The conclusion drawn were the following:
    - The Numerical variables were unevenly distributed.
    - 'Total Time Spent on Website', 'Through Recommendations', 'Lead Source_Wellingak Website', 'Lead Origin_Lead Add Form', 'Lead Source_Google', 'Lead Source_Reference' and 'Last Activity_SMS Sent' had high positive correlation with the target variable
    - 'Do Not Email', Last Activity_Email bounced', 'Lead Source_Olark Chat', 'Last Activity_Olark Chat Conversation' had high negative correlation with the target variable
- Heatmap was plotted for all the variables to check for correlations with each other and redundancies. High correlations were identified and the less important variable(from a business perspective) was removed.

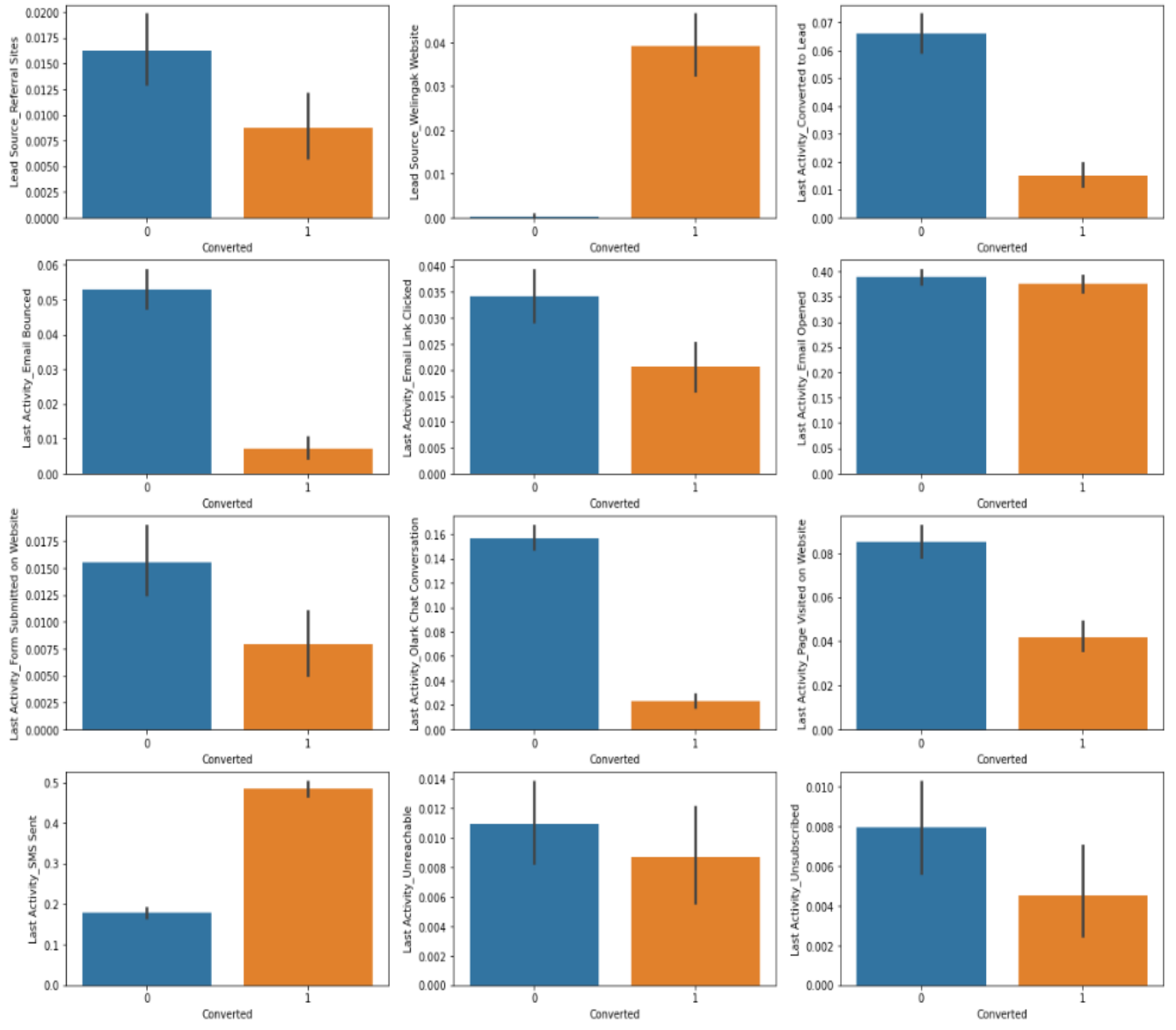


Plot[1,1] and [3,2] having very high negative correlation with target

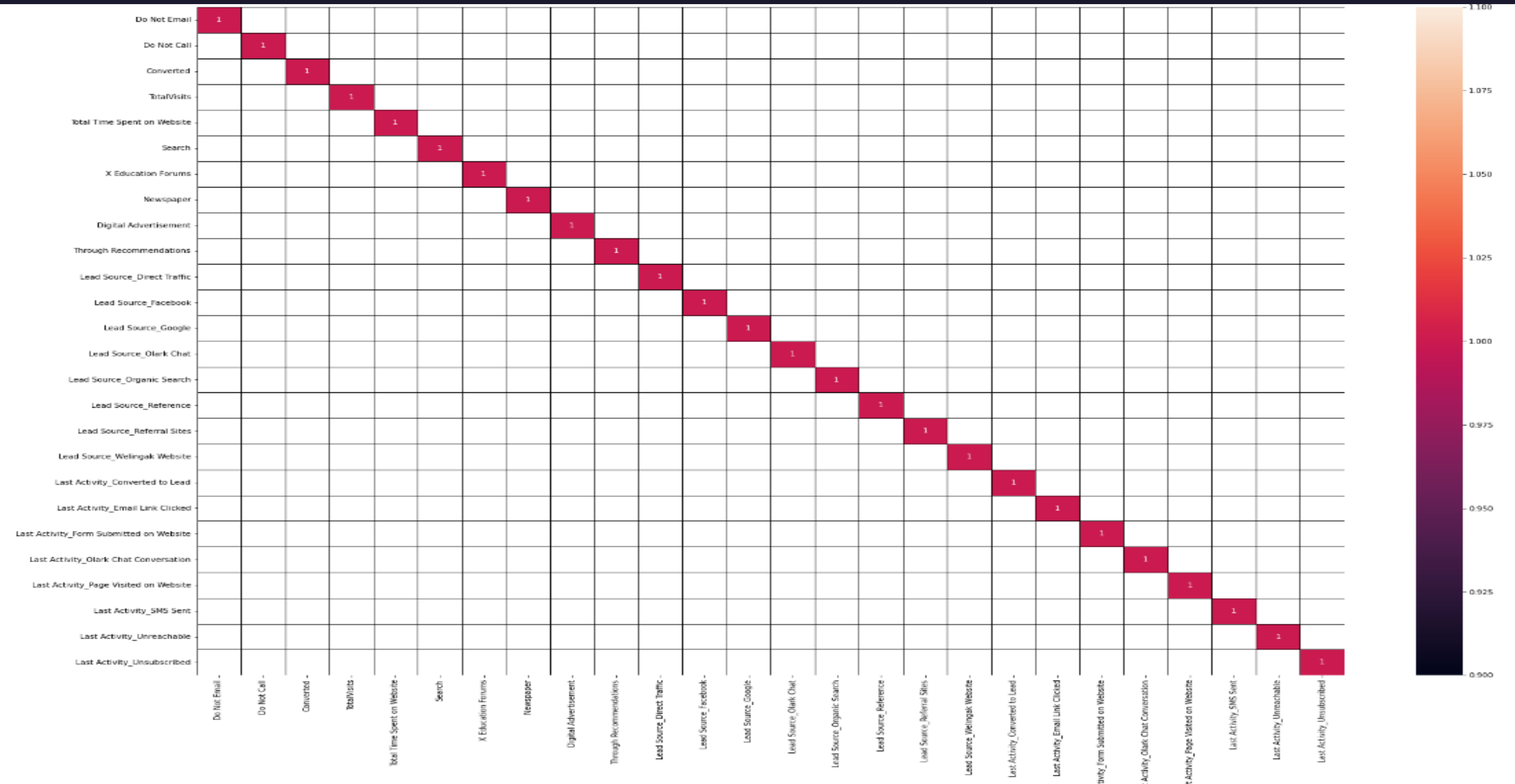
Plot[1,3] and [4,1] having very high positive correlation with target

Plot[1,2], and [4,1] having very high positive correlation with target

Plot[2,1], and [3,2] having very high negative correlation with target



Heatmap after all correlated variables were removed and threshold was taken has <-0.5 and >0.5



Model Building

Feature Selection and Model Creation

- The first step in Model building is to scale the numeric data. We have chosen to Standardize the data in this case to preserve the original distribution of the data.

- Moving to features selection I have used 3 different methods to select the features and then choose the most relevant ones from all three methods. The methods used were: RFE, Univariate Selection (SelectKBest) and Correlation Matrix. Details are given in the notebook.

- Top 15 features were selected from each method and collated in a data frame. After this the selection criteria was the most relevant features chosen by each method by checking their commonality. These features were identified and were used in the model.

	rfe_features	uni_features	cor_features	All matched	rfe & uni	rfe & cor	uni & cor	True count
0	Do Not Email	Do Not Email	Do Not Email	True	True	True	True	4
1	Last Activity_Converted to Lead	Last Activity_Converted to Lead	Last Activity_Converted to Lead	True	True	True	True	4
2	Last Activity_Email Link Clicked	Last Activity_Email Link Clicked	Last Activity_Email Link Clicked	True	True	True	True	4
3	Last Activity_Form Submitted on Website	Last Activity_Form Submitted on Website	Last Activity_Form Submitted on Website	True	True	True	True	4
4	Last Activity_Olark Chat Conversation	Last Activity_Olark Chat Conversation	Last Activity_Olark Chat Conversation	True	True	True	True	4
5	Last Activity_SMS Sent	Last Activity_Page Visited on Website	Last Activity_Page Visited on Website	True	True	True	True	4
6	Lead Source_Direct Traffic	Lead Source_Direct Traffic	Lead Source_Direct Traffic	True	True	True	True	4
7	Lead Source_Google	Lead Source_Olark Chat	Lead Source_Olark Chat	True	True	True	True	4
8	Lead Source_Reference	Lead Source_Referral Sites	Lead Source_Referral Sites	True	True	True	True	4
9	Lead Source_Referral Sites	Lead Source_Welingak Website	Lead Source_Welingak Website	True	True	True	True	4
10	Lead Source_Welingak Website	Total Time Spent on Website	Total Time Spent on Website	True	True	True	True	4
11	Total Time Spent on Website	TotalVisits	TotalVisits	True	True	True	True	4

- A total of four models were trained. At each iteration the feature with the highest p-value was dropped. P-value threshold was considered at 0.05. At the final iteration the VIF was checked and no multicollinearity was found.

Final Model Summary

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6461
Model:	GLM	Df Residuals:	6449
Model Family:	Binomial	Df Model:	11
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-3003.6
Date:	Sat, 21 Jan 2023	Deviance:	6007.2
Time:	17:10:14	Pearson chi2:	6.65e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.5509	0.065	-8.520	0.000	-0.678	-0.424
Do Not Email	-1.3852	0.147	-9.444	0.000	-1.673	-1.098
Last Activity_Converted to Lead	-1.2088	0.195	-6.194	0.000	-1.591	-0.826
Last Activity_Email Link Clicked	-0.5210	0.199	-2.621	0.009	-0.911	-0.131
Last Activity_Form Submitted on Website	-0.6685	0.301	-2.223	0.026	-1.258	-0.079
Last Activity_Olark Chat Conversation	-1.2631	0.150	-8.425	0.000	-1.557	-0.969
Last Activity_SMS Sent	1.1421	0.070	16.202	0.000	1.004	1.280
Lead Source_Direct Traffic	-0.6939	0.086	-8.069	0.000	-0.862	-0.525
Lead Source_Google	-0.3554	0.081	-4.379	0.000	-0.514	-0.196
Lead Source_Reference	3.3510	0.207	16.173	0.000	2.945	3.757
Lead Source_Welingak Website	5.6737	1.010	5.616	0.000	3.694	7.654
Total Time Spent on Website	1.0131	0.035	29.055	0.000	0.945	1.081

	features	VIF
5	Last Activity_SMS Sent	1.36
7	Lead Source_Google	1.30
6	Lead Source_Direct Traffic	1.27
10	Total Time Spent on Website	1.13
0	Do Not Email	1.09
1	Last Activity_Converted to Lead	1.07
4	Last Activity_Olark Chat Conversation	1.07
8	Lead Source_Reference	1.06
2	Last Activity_Email Link Clicked	1.03
3	Last Activity_Form Submitted on Website	1.03
9	Lead Source_Welingak Website	1.03

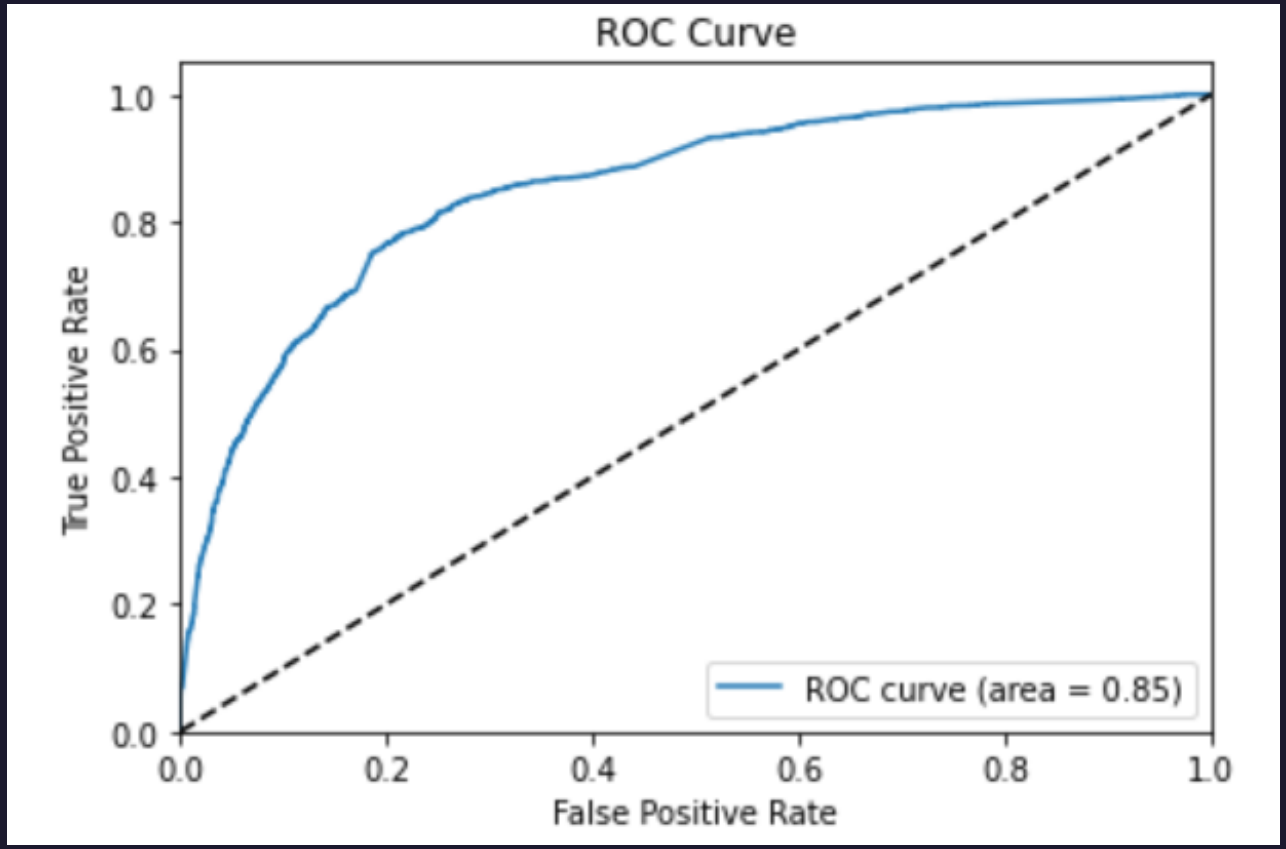
VIF data frame

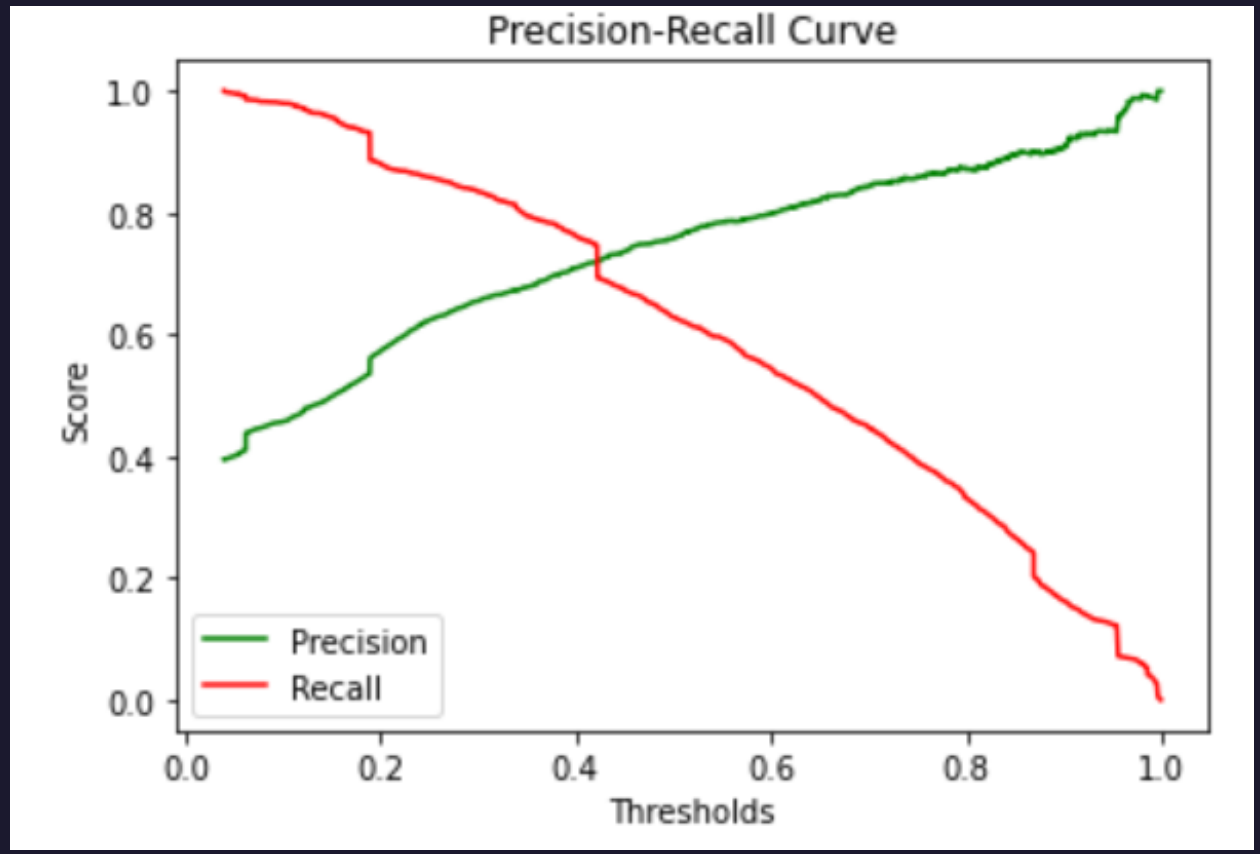
Model Evaluation

ROC Curve, Precision and Recall

- Initial threshold value of 0.5 was chosen and metric calculated on the training set predictions were:
  - Precision = 0.758
  - Recall = 0.628
- As the recall is very low an optimal value of threshold has to be chosen. This was done by plotting the ROC curve and the Precision_Recall curve. Based on these two curves different threshold values were tried. 0.3 was chosen as optimal since it gave a high recall without sacrificing too much precision and accuracy. As highlighted earlier, recall is the more relevant metrics for our business case.
- Predictions were made on the test set using 0.3 as threshold and the metrics obtained are given below
- As we can see, the model is performing very well on unseen data
- The target conversion rate has also been achieved by having a recall score of >80%

	Training Metrics	Test Metrics
Accuracy	0.765	0.779
Precision	0.655	0.662
Recall	0.835	0.853


ROC Curve
True Positive Rate
False Positive Rate
ROC curve (area = 0.85)
0.0
0.2
0.4
0.6
0.8
1.0


Precision-Recall Curve
Score
Thresholds
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0

- Lead score was also assigned to the test set based on the probabilities and an additional binary recommendation column was added
- Based on the scores and recommendation operators can now easily identify which leads are the most promising and pursue them. This wills save time, money and improve the conversion rate

Prospect ID	Original Converted	Lead Score	Recommendation
52a20b92-5a45-499e-8631-d958713adab1	No	4.81	No
891e04ef-6ac2-4d6e-a308-f1d2517daee5	Yes	48.52	Yes
ae107016-c7dc-49a2-bd3d-6daa1ad4de1a	Yes	71.23	Yes
0c73eba1-4963-4a03-9f9d-b4afb1dacef5	No	18.88	No
81a23cd8-e8ca-4484-afbd-02ed7366cee2	No	50.55	Yes
...	...	...	...
d1af0b05-e77f-4111-811c-7c7abe1919e2	No	16.91	No
c9bbd996-a573-453d-b1f7-a52f3971ac7e	No	14.34	No
a0f34735-cbbf-49f1-aac1-fbed6695ef0f	Yes	77.15	Yes
623c7e43-c80c-44ef-a640-2815873ab838	Yes	63.72	Yes
f4f94e42-16f3-47b5-946a-11b7909421f0	No	62.90	Yes

Interpretation and Recommendations

- By building this solution we have identified some of the key factors that the business can focus on to generate more relevant leads. This way resources can be allocated to the most important areas. So based on our EDA and then fine tuned by the model the business can focus on these 5 driver variables:
    - Lead Source_Wellingak Website, Lead Source_Reference, Last Activity_SMS Sent, Total Time Spent on Website and Do Not Email
    - We can infer that if the lead originated from Wellingak Website or if it came from a reference then it is more likely to get converted
    - If the customer has sent an SMS as their last activity then they are more likely to be converted
    - The mode time a person spends on the website, the more likely they are to be converted
    - If the person has NOT opted to be notified by email, meaning they are willing o accept email communication, they are more likely to be converted
- The business should focus on monitoring these variables and allocate resources in marketing on these channels to maximize the conversion rate. This can be done by:
    - Spreading awareness about the website and adding hyperlinks to the website from other webpages and ads
    - Enhance or create a better referral system so more leads can be generated through reference
    - Communication through SMS or Email can lead to better lead conversion, so these channels should be prioritised

Thank You

Sharod Dey