Lead Scoring Case Study

Business Approach for X Education

- By:
- Abhishek Mehandiratta
- Sudhanshu Raj
- Harshal Kl

Objective

- Helping X Education identify promising leads from all the leads generated for their product. This will help them focus more on potential customers, hence increasing revenue and employee productivity.

Desired outcome and methodology

- Building a Logistic regression model on the leads data provided by X Education.
- This model assigns a score to the leads such that a higher score means higher conversion chance.

- Target leads conversion score is around 80%.

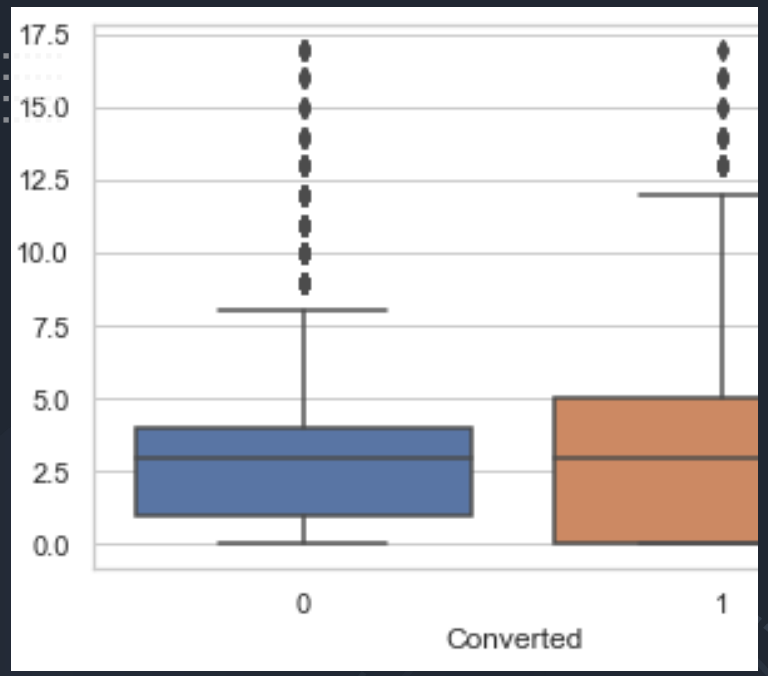
Desired outcome and methodology

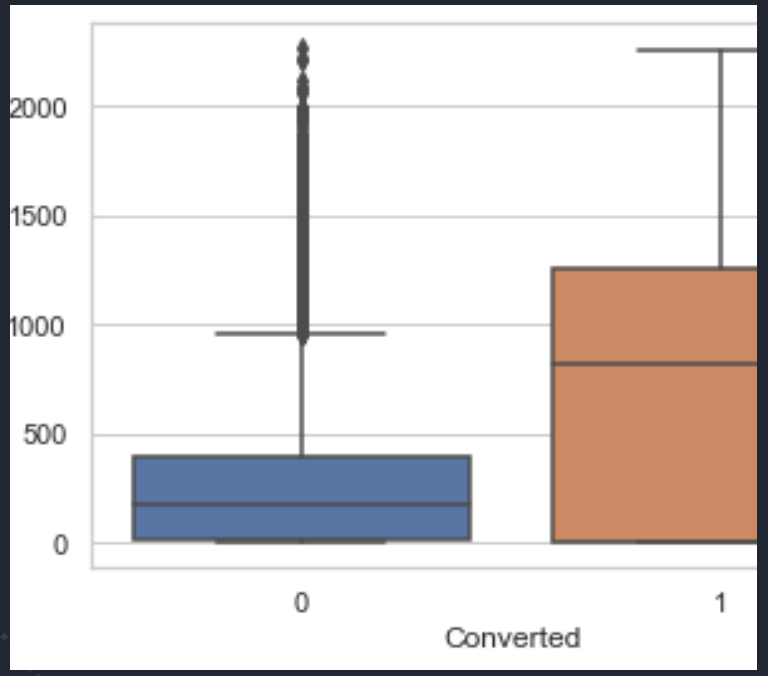
- Steps involved:
- Data cleaning, manipulation and transformation:
    - Handling missing values (dropping columns with too many missing values and imputing some columns to not lose out on too much information).
    - Handling outliers in the numerical columns (dropping values less than 1st percentile and greater than 99th percentile if the column has outliers).
- EDA
    - Univariate analysis:
        - On numerical features (identified outliers and dropped them).
        - On categorical variables (identified columns with little to no variance and dropping them because they don't add any value to our analysis).
    - Bivariate analysis:
        - Identifying correlations between different variables.
- Feature Scaling using Min/max scaling.
- Applying OHE (one hot encoding) to categorical variables and converting them to dummy variables.

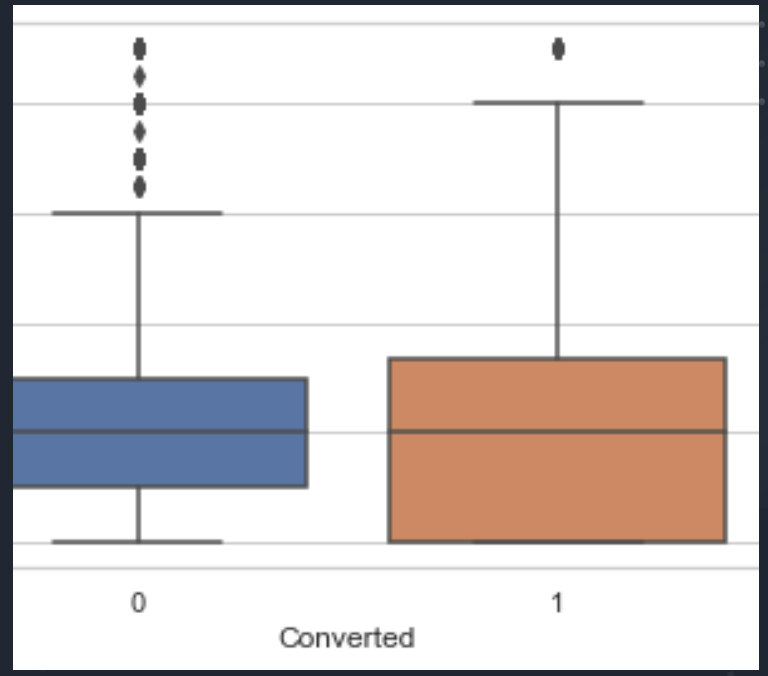
Desired outcome and methodology

- Steps involved:
- Model Building:
    - Selecting initial feature set using RFE
    - Using Statsmodel to identify statistically insignificant variables and dropping them, thereby creating a better model.
    - Validation of model using metrics such as accuracy, specificity and sensitivity.
    - Fine-tuning the model by selecting an optimal cutoff using the ROC curve and maintaining a balance between accuracy, specificity and sensitivity across different cutoffs and selecting the best one.
- Model conclusions, interpretation and recommendation

EDA and some insights

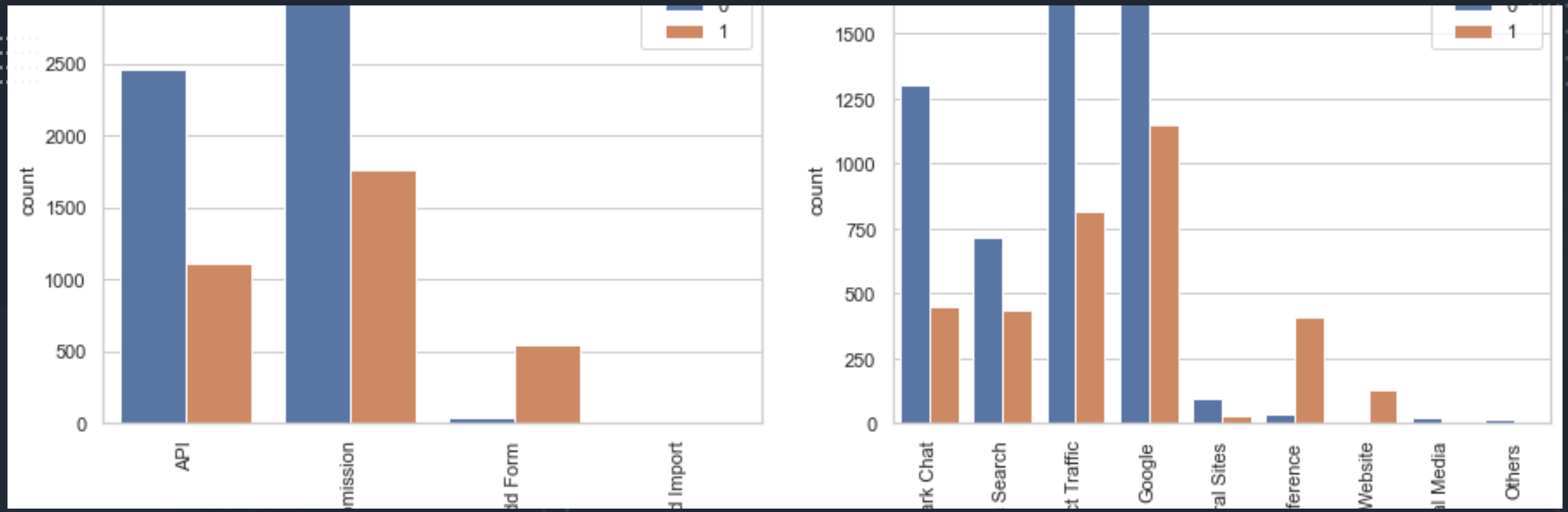






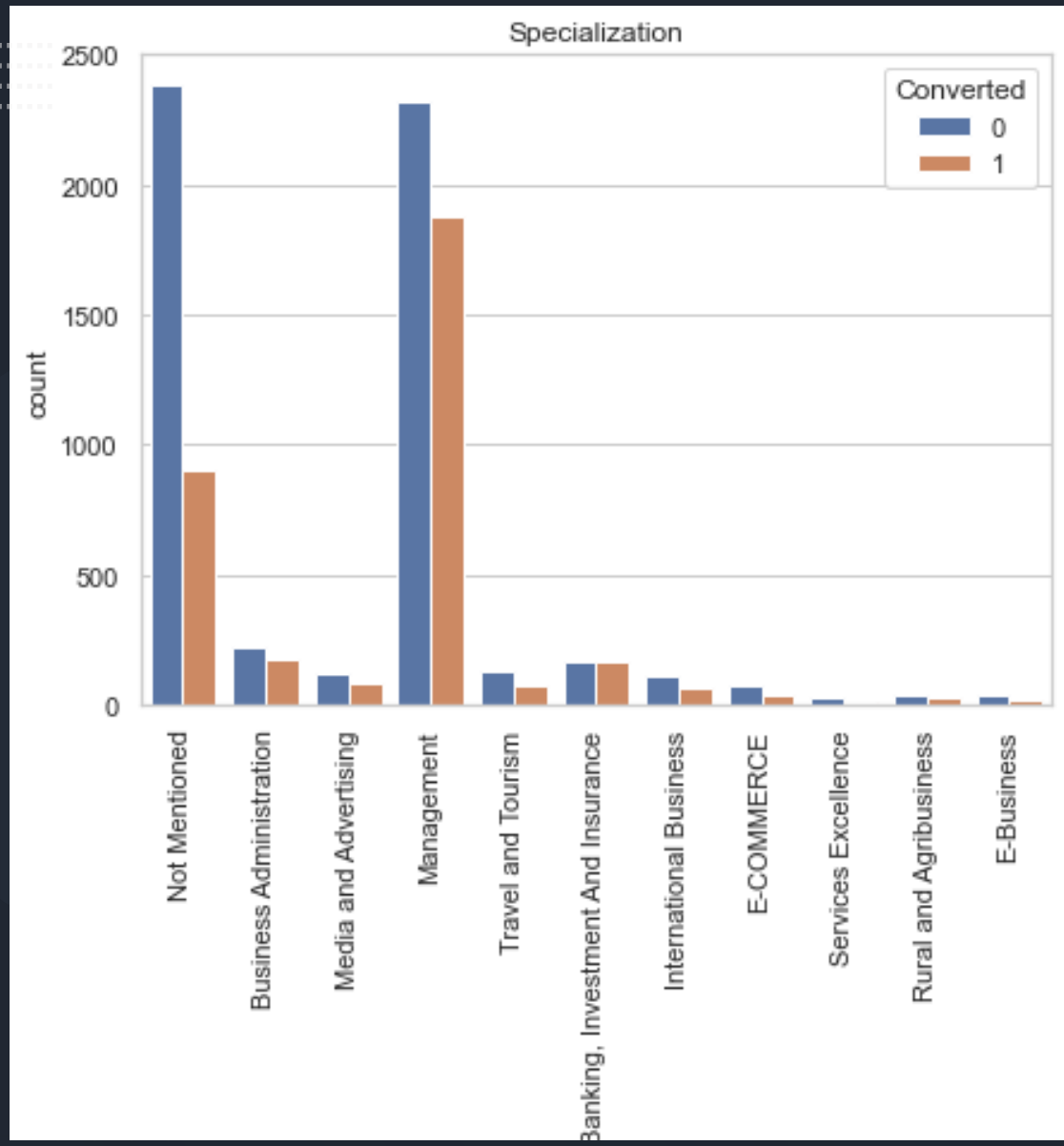
Numerical Variables

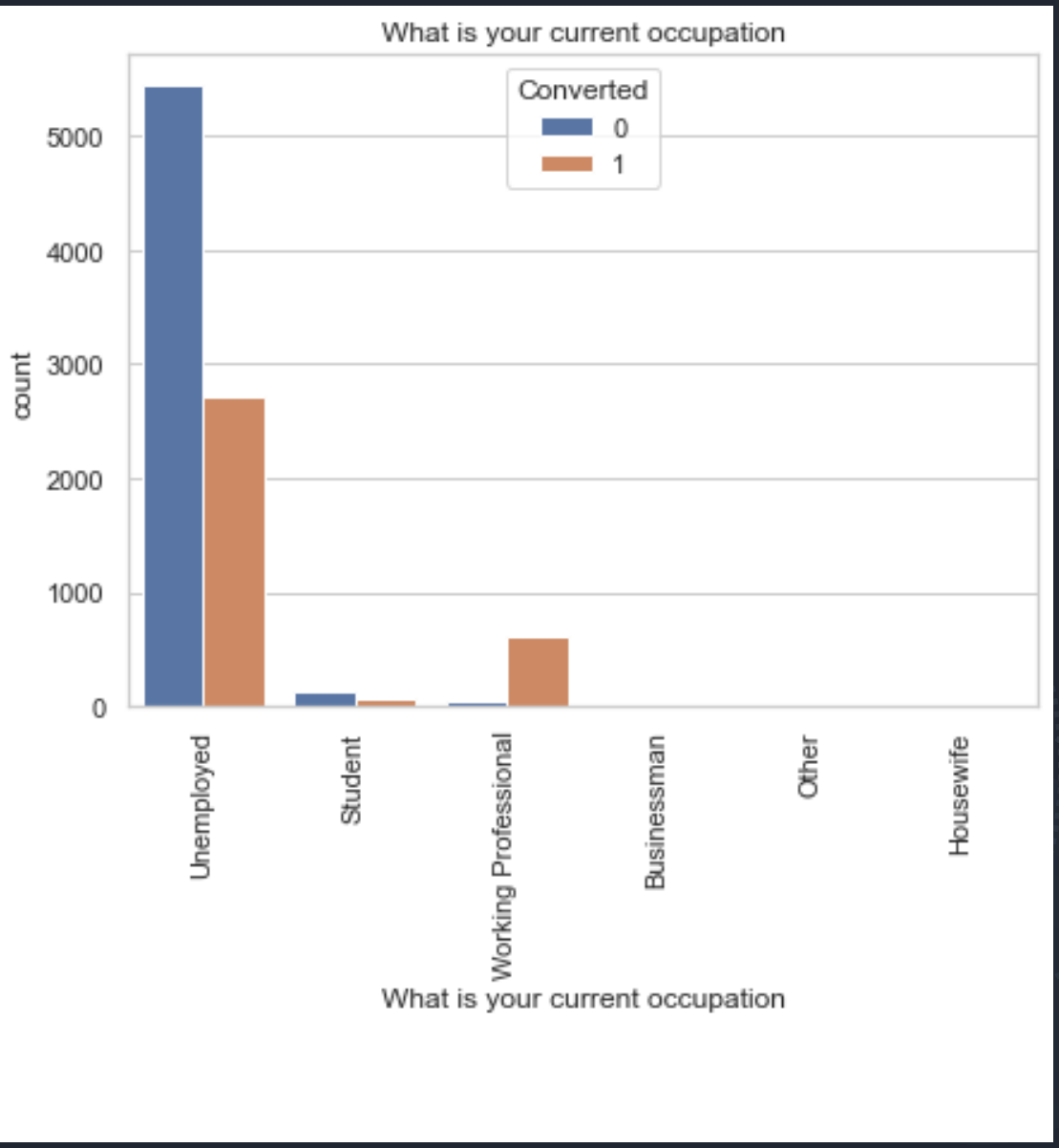
- Converted leads spend significantly more time on the
- website than the non-converted ones.



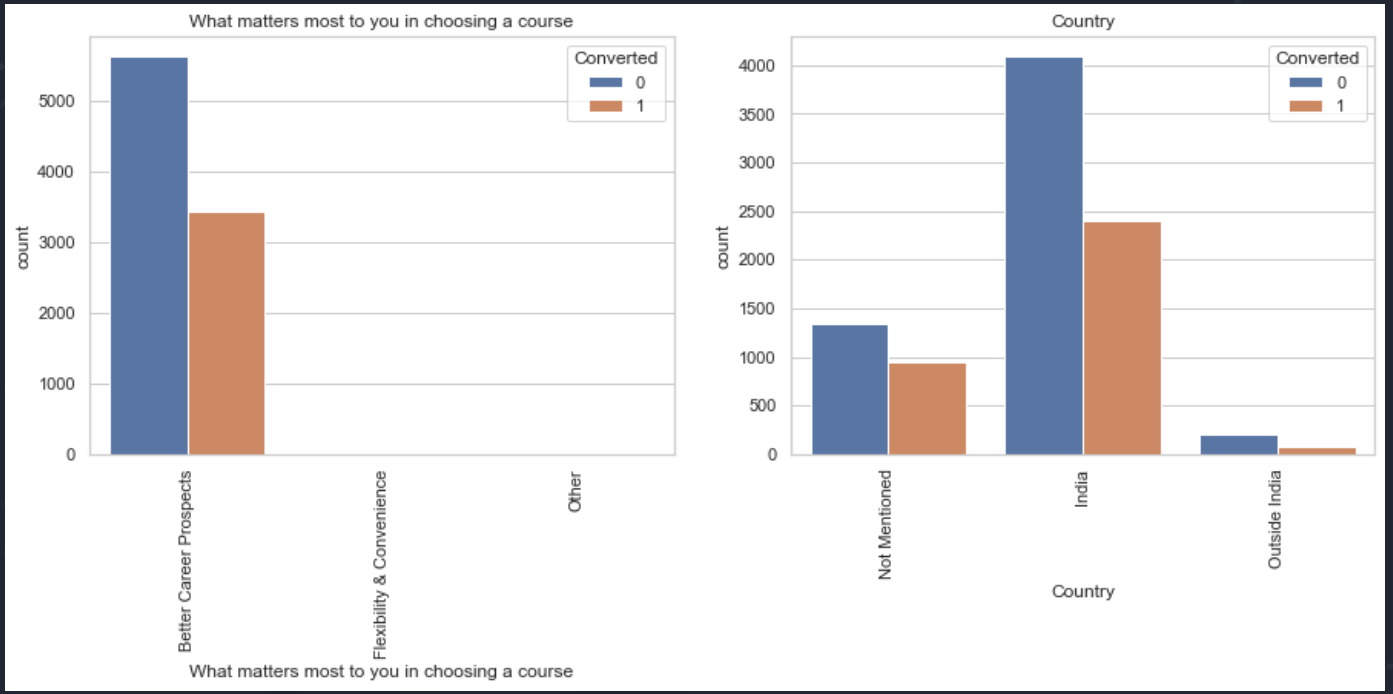
Categorical Variables

- Majority of the Leads were Identified at the "Landing Page Submission”.
- "Olark Chat", "Direct Trafic" and "Google" were major Lead sources and
- amongst them Google saw highest lead conversion.


Specialization
Converted
0
1
count
2500
2000
1500
1000
500
0
Not Mentioned
Business Administration
Media and Advertising
Management
Travel and Tourism
Banking, Investment And Insurance
International Business
E-COMMERCE
Services Excellence
Rural and Agribusiness
E-Business


What is your current occupation
Converted
0
1
count
5000
4000
3000
2000
1000
0
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation

- This plot shows a particular reason why people are choosing a course.
- We see that most of the leads are from India.



Model Evaluation

We see that all the features selected have 0 p-values,

which means all these features are statistically

significant.

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6246
Model:	GLM	**Df Residuals:**	6234
Model Family:	Binomial	**Df Model:**	11
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1902.6
Date:	Mon, 14 Nov 2022	**Deviance:**	3805.2
Time:	23:56:43	**Pearson chi2:**	6.67e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.5128
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.3556	0.094	-24.936	0.000	-2.541	-2.170
TotalVisits	1.8763	0.317	5.917	0.000	1.255	2.498
Total Time Spent on Website	3.7865	0.181	20.882	0.000	3.431	4.142
Page Views Per Visit	-3.3847	0.310	-10.907	0.000	-3.993	-2.777
Lead Source_Reference	2.6386	0.256	10.323	0.000	2.138	3.140
Lead Source_Welingak Website	5.5125	0.733	7.518	0.000	4.075	6.950
Do Not Email_Yes	-1.5408	0.214	-7.216	0.000	-1.959	-1.122
Last Activity_Olark Chat Conversation	-1.0953	0.192	-5.700	0.000	-1.472	-0.719
What is your current occupation_Working Professional	1.2509	0.237	5.282	0.000	0.787	1.715
Tags_Other_Tags	0.7066	0.093	7.594	0.000	0.524	0.889
Tags_Will revert after reading the email	4.6629	0.177	26.288	0.000	4.315	5.011
Last Notable Activity_SMS Sent	1.6917	0.099	17.150	0.000	1.498	1.885

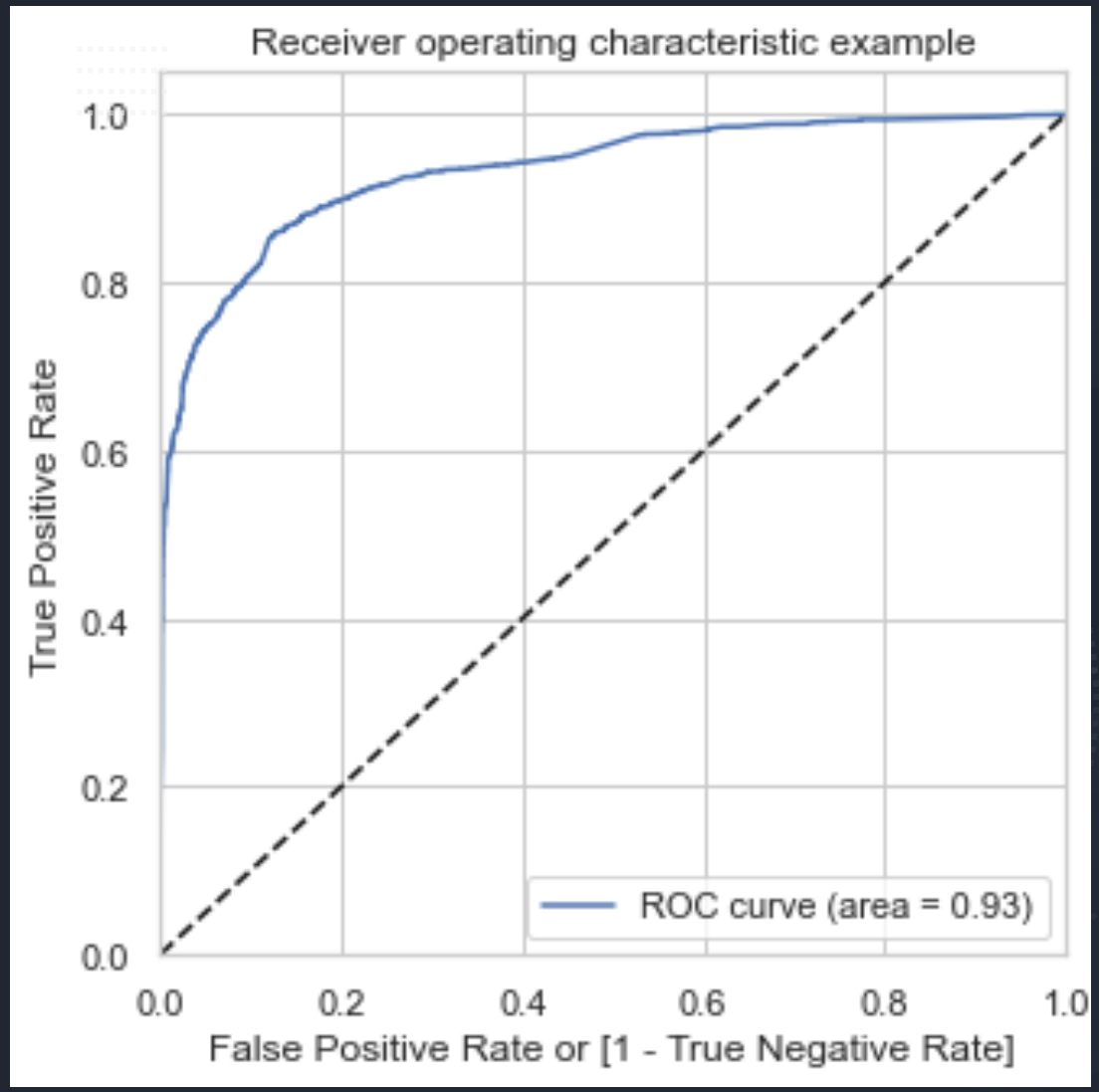
Multicollinearity

- We can see that the VIF values are all less than 5.
- Hence there is no multicollinearity among the final
- features.

	Features	VIF
2	Page Views Per Visit	4.81
0	TotalVisits	4.58
1	Total Time Spent on Website	2.23
9	Tags_Will revert after reading the email	1.93
10	Last Notable Activity_SMS Sent	1.45
8	Tags_Other_Tags	1.32
7	What is your current occupation_Working Profes...	1.29
3	Lead Source_Reference	1.22
5	Do Not Email_Yes	1.07
6	Last Activity_Olark Chat Conversation	1.03
4	Lead Source_Welingak Website	1.02

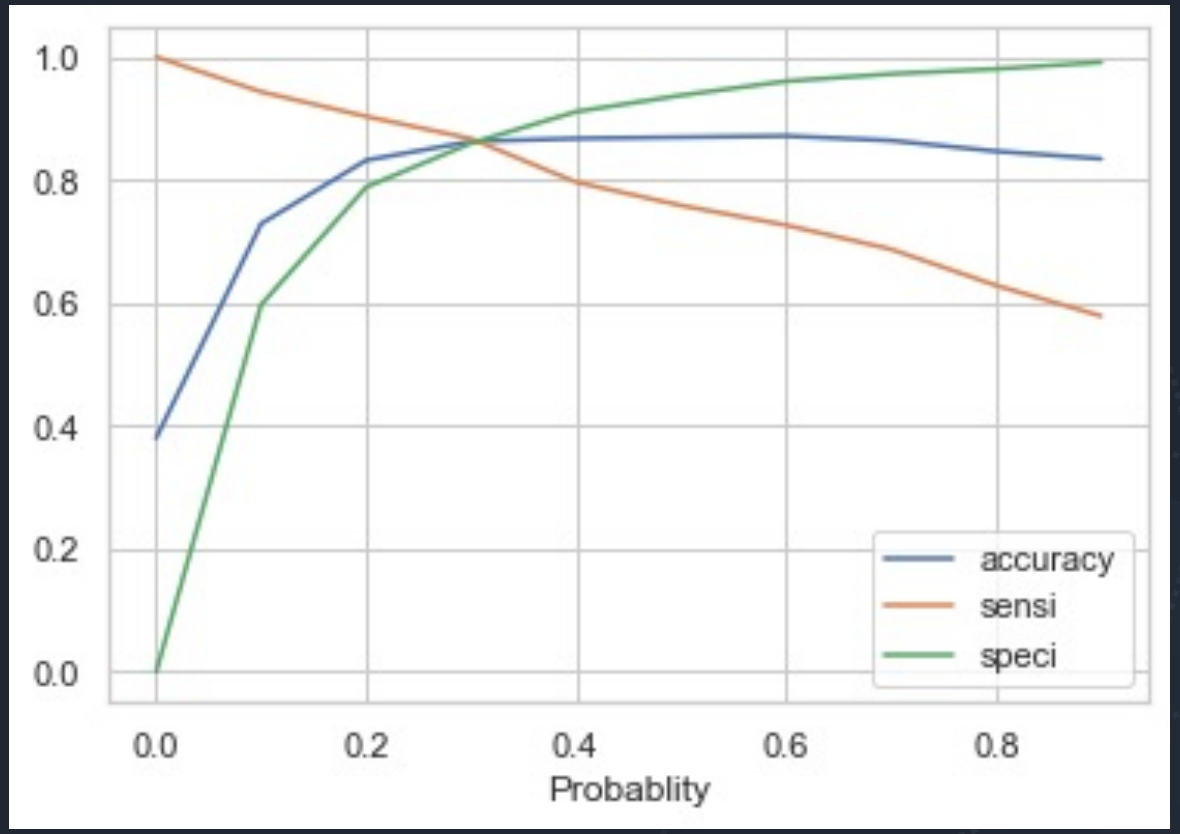
ROC Curve

- We can see that the area under ROC
- is 0.93, which means we have a good
- predictive model.



Optimal threshold

- We can see that the optimal threshold is 0.3.



```
In [146]: confusion3 = metrics.confusion_matrix(y_pred_final.Converted, y_pred_final.Final_Prediction )
          confusion3

Out[146]: array([[1435,  249],
                 [ 133,  861]])
```

Final Confusion Matrix

Metrics on train and test set

Train set:

- Accuracy: 86%
- Sensitivity: 86%
- Specificity: 85%

```
# Check the overall accuracy

metrics.accuracy_score(y_train_pred_final.Converted, y_train_pred_final.F
```

```
0.8624719820685238
```

```
# Creating confusion matrix

confusion2 = metrics.confusion_matrix(y_train_pred_final.Converted, y_tra
confusion2
```

```
array([[3328,  543],
       [ 316, 2059]])
```

```
# Substituting the value of true positive
TP = confusion2[1,1]
# Substituting the value of true negatives
TN = confusion2[0,0]
# Substituting the value of false positives
FP = confusion2[0,1]
# Substituting the value of false negatives
FN = confusion2[1,0]
```

```
# Calculating the sensitivity

TP/(TP+FN)
```

```
0.8669473684210526
```

```
# Calculating the specificity

TN/(TN+FP)
```

```
0.8597261689485921
```

Test set:

- Accuracy: 86%
- Sensitivity: 87%
- Specificity: 86%

- Precision: 78%
- Recall: 87%

```
# Let's check the overall accuracy.
round(100 * (metrics.accuracy_score(y_pred_final.Converted, y_pred_final.Final_Prediction)))
```

```
86
```

```
confusion3 = metrics.confusion_matrix(y_pred_final.Converted, y_pred_final.Final_Prediction )
confusion3
```

```
array([[1435,  249],
       [ 133,  861]])
```

```
TP = confusion2[1,1] # true positive
TN = confusion2[0,0] # true negatives
FP = confusion2[0,1] # false positives
FN = confusion2[1,0] # false negatives
```

```
# Let's see the sensitivity of our logistic regression model
round(100 * (TP / float(TP+FN)))
```

```
87
```

```
# Let us calculate specificity
round(100 * (TN / float(TN+FP)))
```

```
86
```

```
from sklearn.metrics import precision_score, recall_score
```

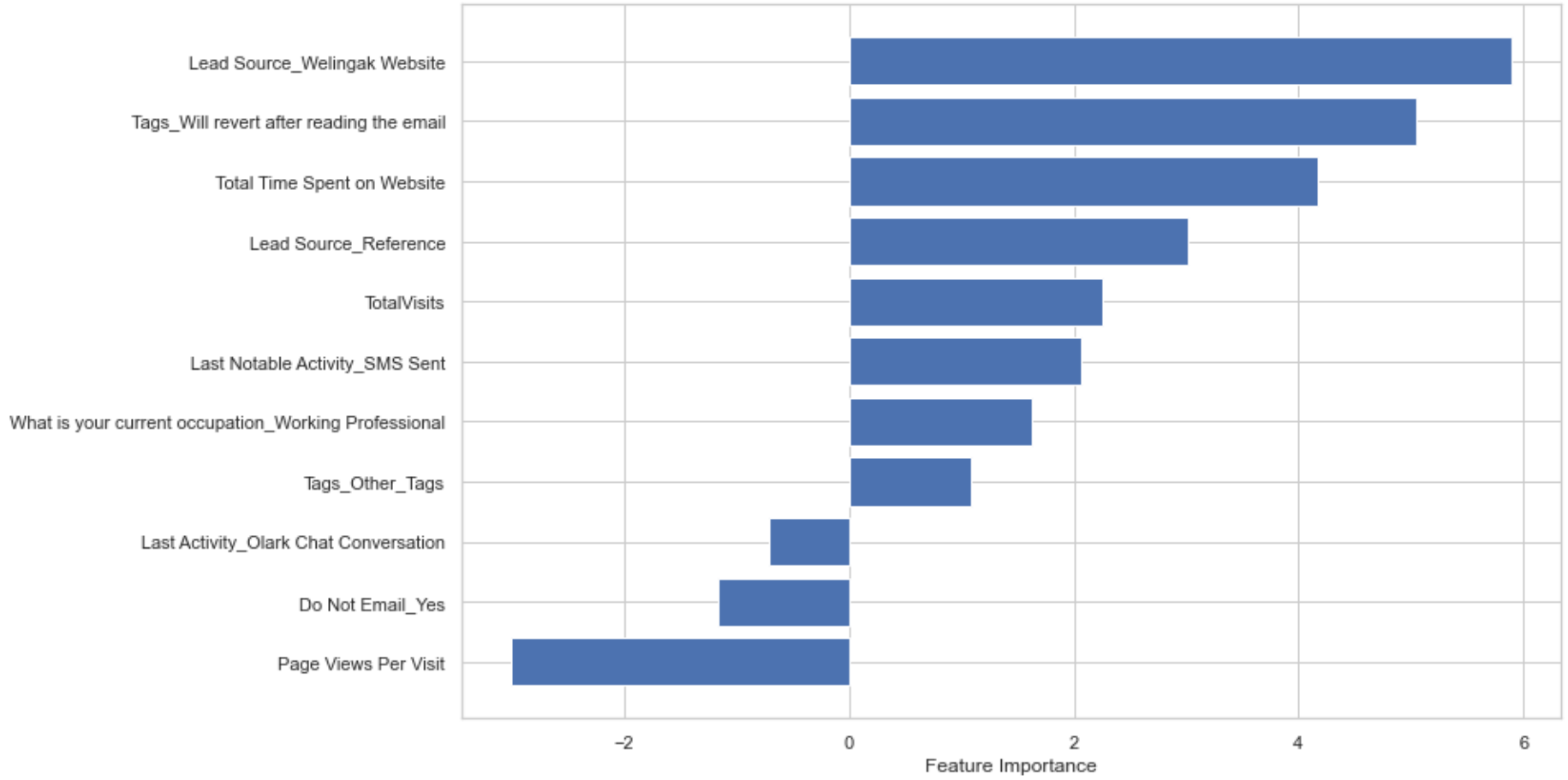
```
round(100 * (precision_score(y_pred_final.Converted , y_pred_final.Final_Prediction)))
```

```
78
```

```
round(100 * (recall_score(y_pred_final.Converted, y_pred_final.Final_Prediction)))
```

```
87
```

Feature importance



Conclusions & Recommendations

Based on the feature importance graph in the previous slide

- Leads coming from Welingak website are also potentially most likely to be converted.
- Leads Tagged with “Will revert after reading email” have the most impact on the conversion rate.
- Total time spent on website is directly proportional to the probability of conversion.
- Based on business needs, the probability threshold value can be changed for identifying potential leads.

Thank you