Lead Scoring for X Education

Group Project by Lakshmi Sagar SP and Deepika L

Objective

An education company named X Education sells online courses to industry professionals. On any given day, many professionals interested in the courses land on their website, and other avenues, etc. and browse for courses. Once they fill up a form with their phone number and email id, a lead is generated. Currently, the lead conversion rate is about 30%.

Our objective is to build a logistic regression model to assign scores between 0 to 100 to the leads procured, so that the marketing team can focus on the hotter leads, increasing their conversion rate and overall profits.

The model also needs to adjust to the company's requirements if they were to change in the future.

Approach

Columns with more than 40-50% missing values and unwanted columns have been dropped. Extra rows, if any, are deleted.

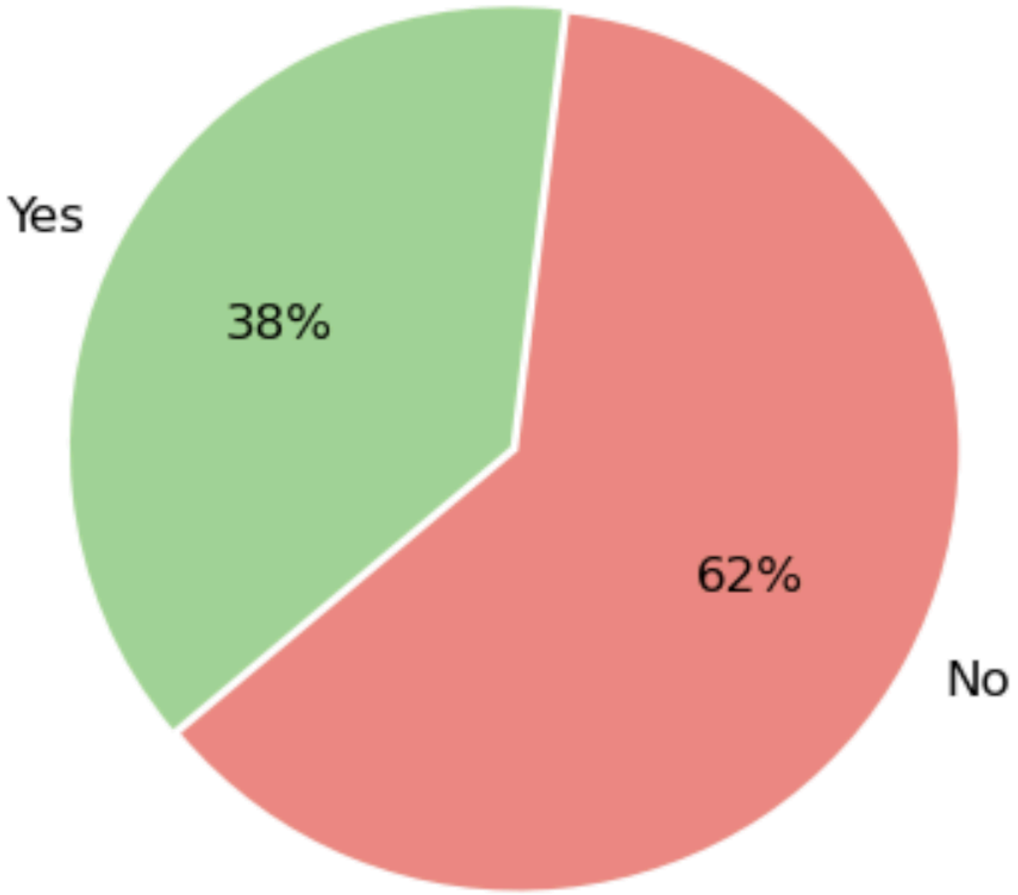
Missing values in Categorical columns have been replaced with the column's mode in general or renamed as accordingly.

Missing values in numerical columns have been imputed with median, mean, or mode or left as missing depending on the column and the business meaning.

Outliers have been treated accordingly. Absurd/impossible values have been dropped.

All the columns have been converted to the right format and standardized for ease of analysis and better visualization.

Data Imbalance: Converted Vs. Non Converted

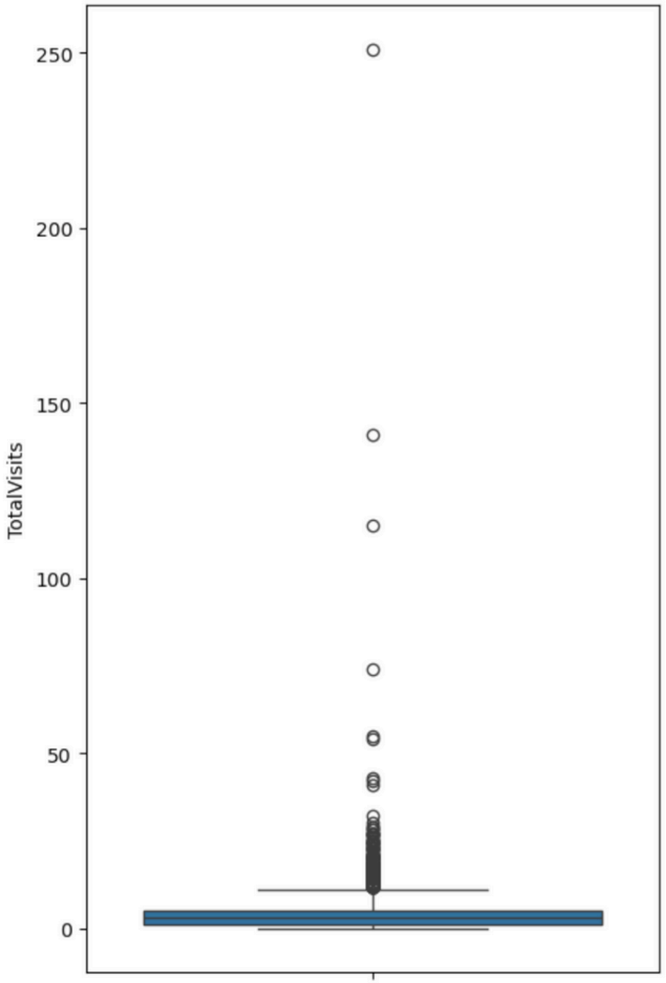


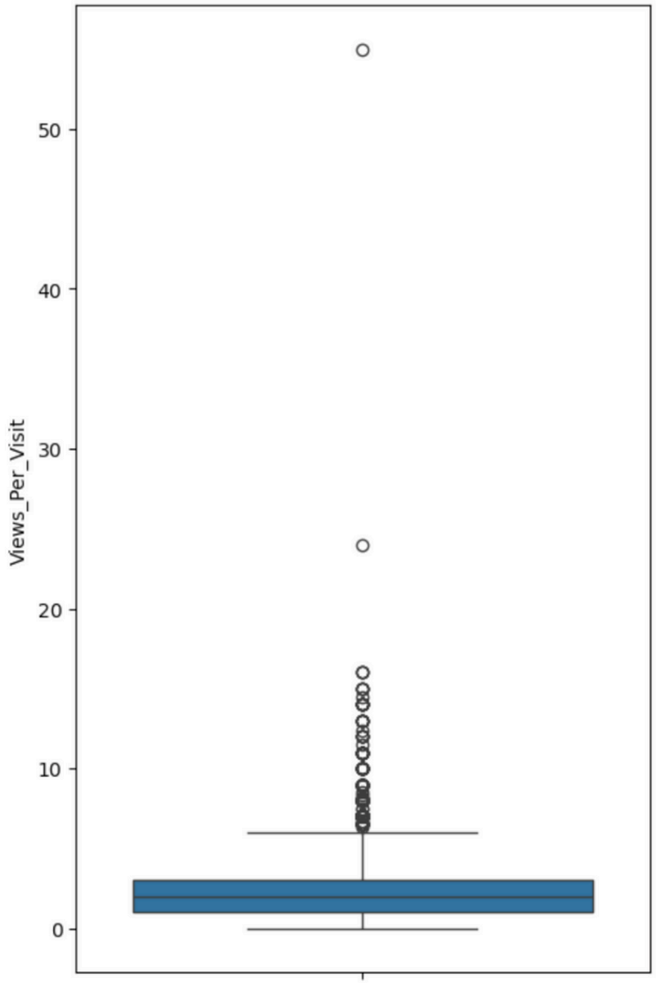
There is a mild imbalance of not-converted leads in the data.

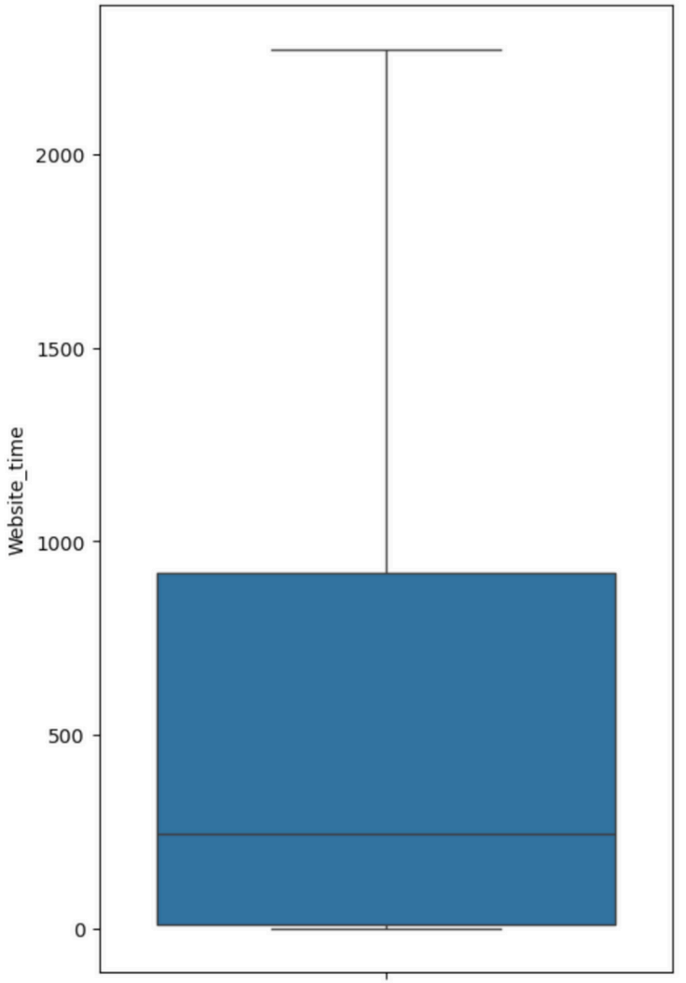
For every 5 converts, there are 8 leads who do not convert.

Summary Statistics

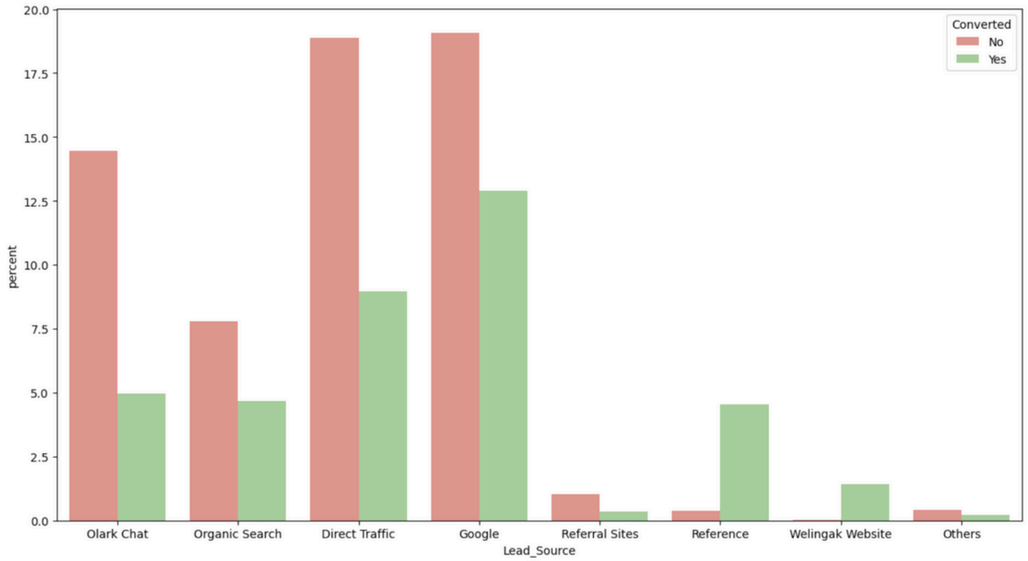
Outliers in the data have been capped to 99th percentile to avoid spam/bot action and rare occurrences.



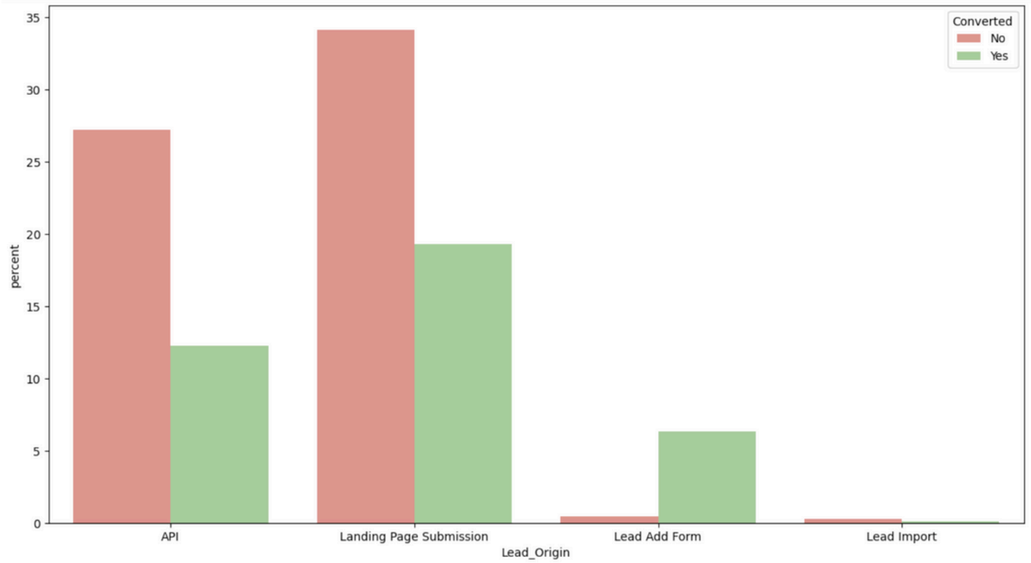


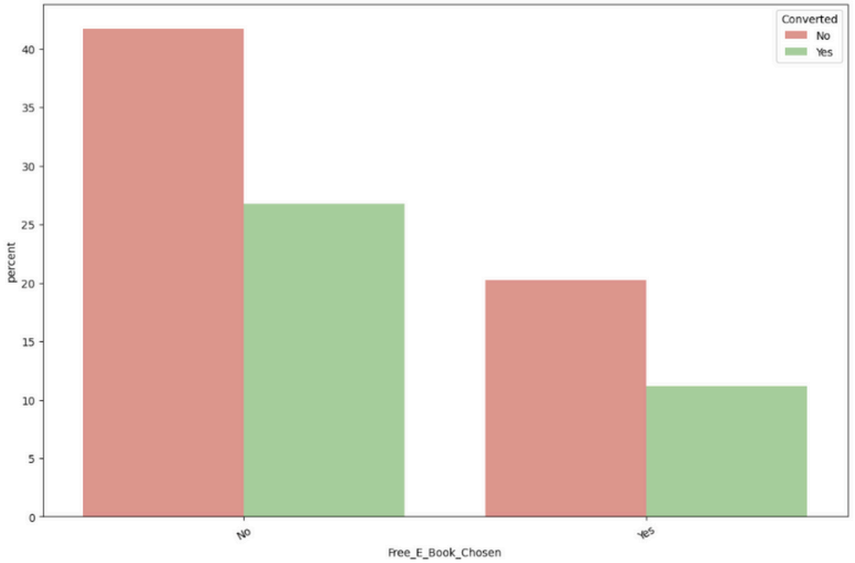


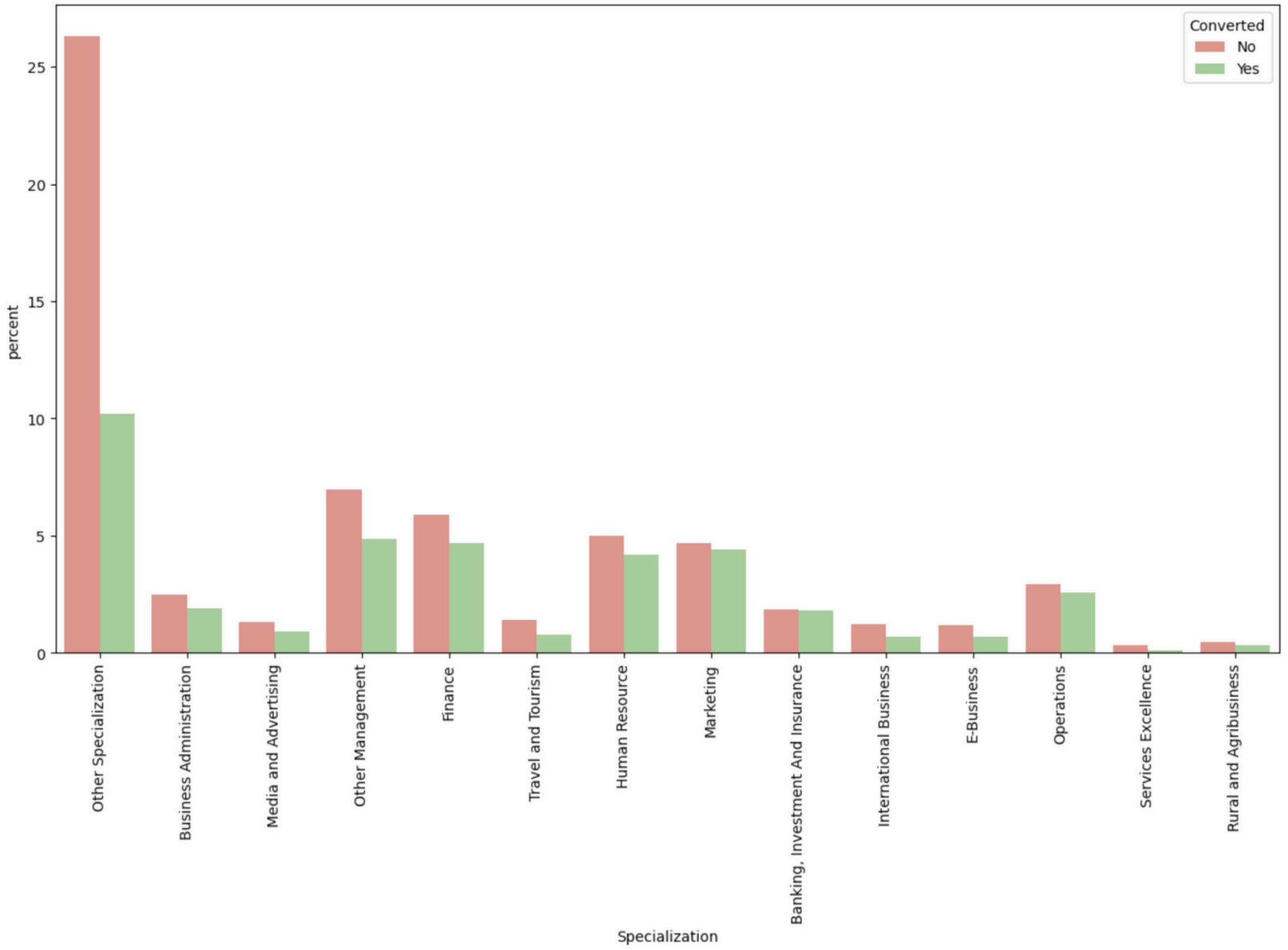
EDA Findings



- Major leads originated from API or landing page submissions.
- Google, direct traffic and organic search are generating most of the traffic and leads.
- Olark chat conversion can be better by focusing on the chat experience.
- Good distribution between leads who chose to get the E-book. The company can add more value to the book to gain trust.

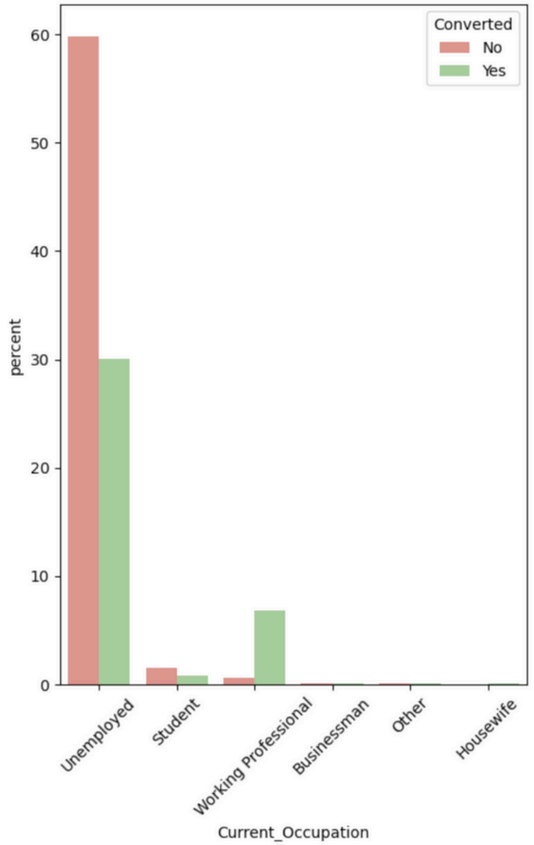


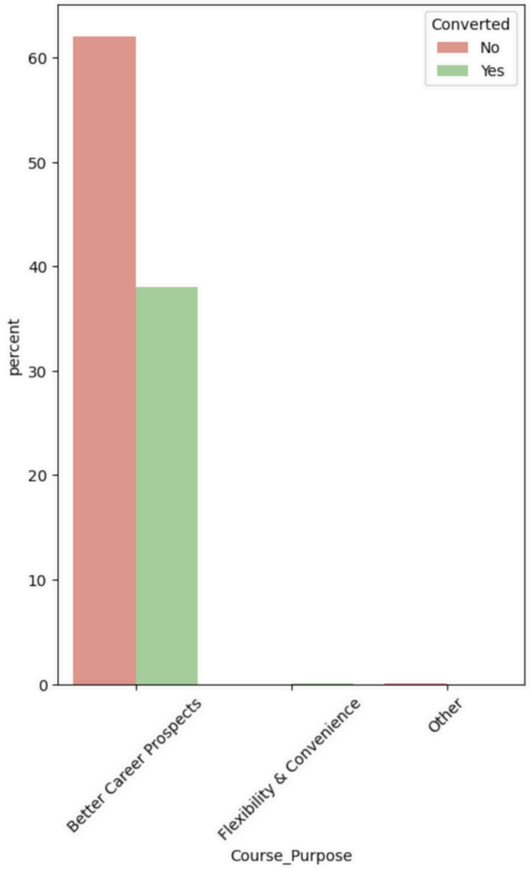


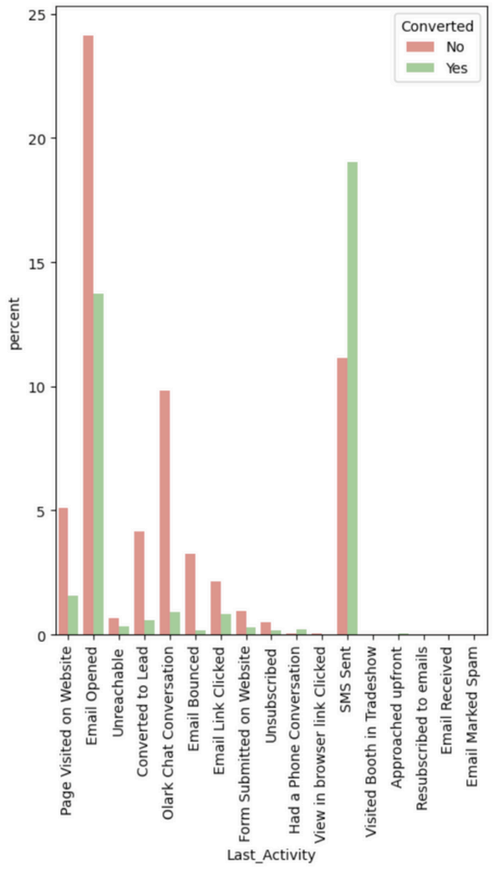


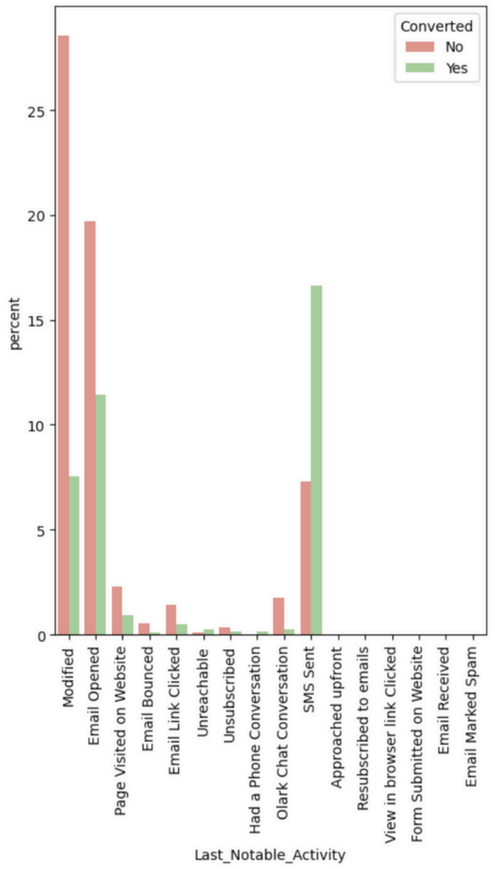
- Most of the data was not specified.
- This suggests these users are either students in undergrad or below (non-specialisation), belong to Tech,
- Or do not have formal education, or Houseviwes, etc.

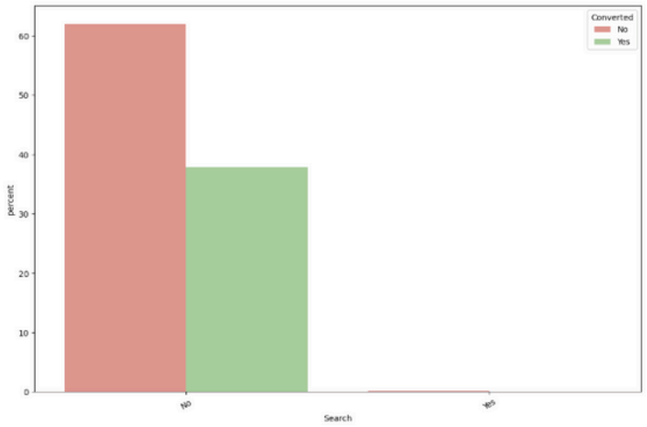
- Unemployed leads were a majority and “Better Career Prospects” was chosen by 96% of them.
- This shows that there is a demand for such courses and imparting employable skills.
- The sales team can be trained to convince the users that their courses can help them build better careers.
- If company can deliver on their promise by impart high quality knowledge and helping them build better careers, this will help gain trust and improve brand loyalty. “MORE REFERRALS”

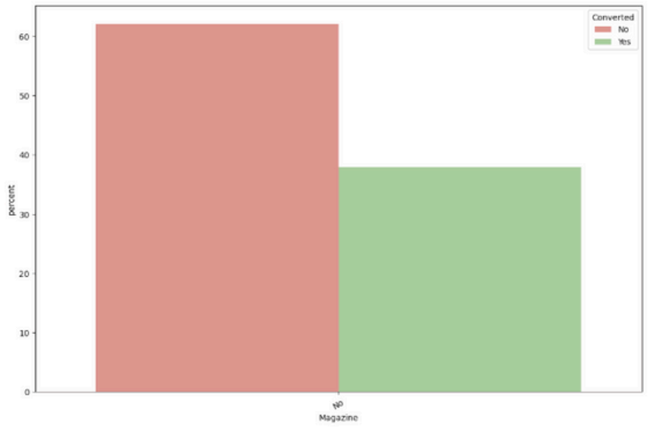


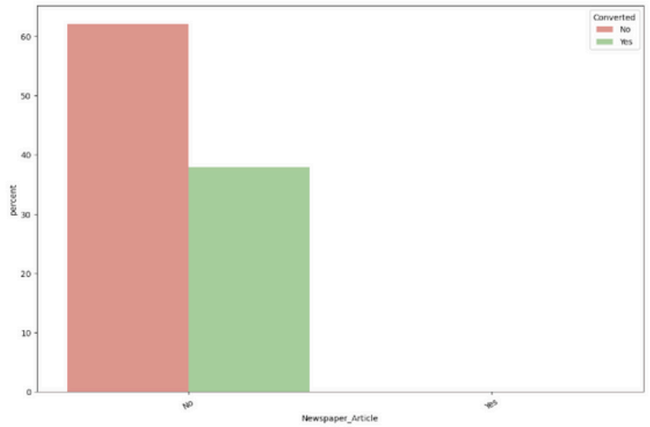


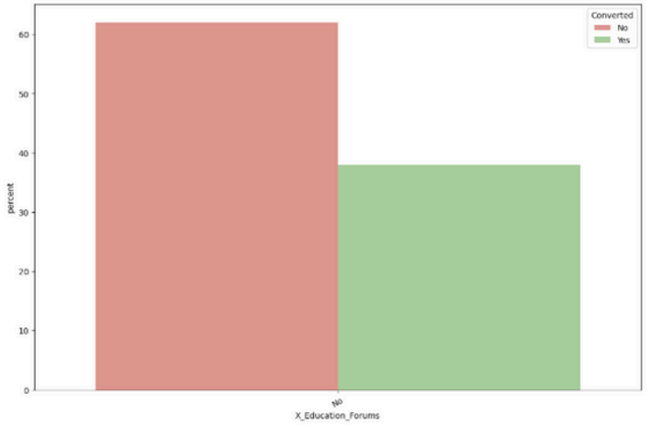


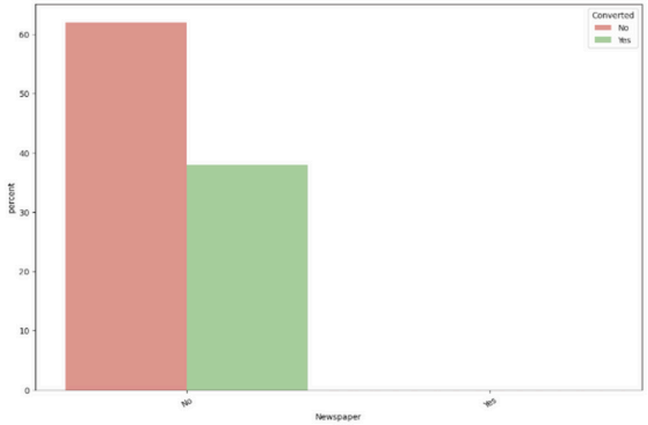


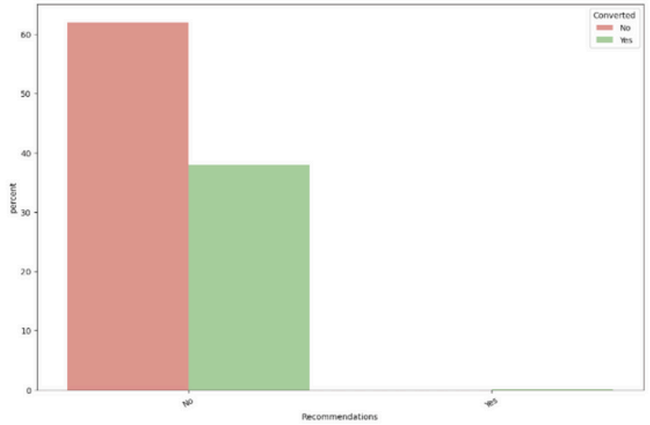


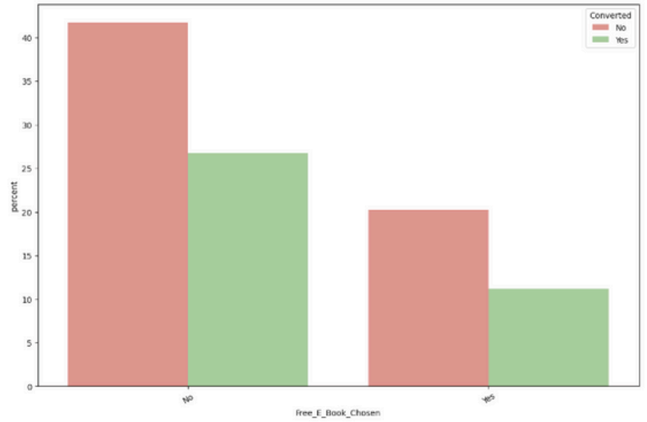












- Marketing efforts did not provide any valuable leads. The columns were either 100% ‘no’ or 95%+ biased towards ‘no’.
- The company can improve their marketing efforts and invest in delivering good messages so that it gets a good ROI on its ad spend.
- ~35% of the leads chose to get a free-ebook but a majority failed to convert. Company needs to improve the value add with the e-book.

Data Preperation

Variables that were not user-generated like ‘Tags’, ‘Last Activity’, etc. did not add value to this model so they were dropped.

This would give a model with features that can score a lead before any communication attempt is made which is the objective.

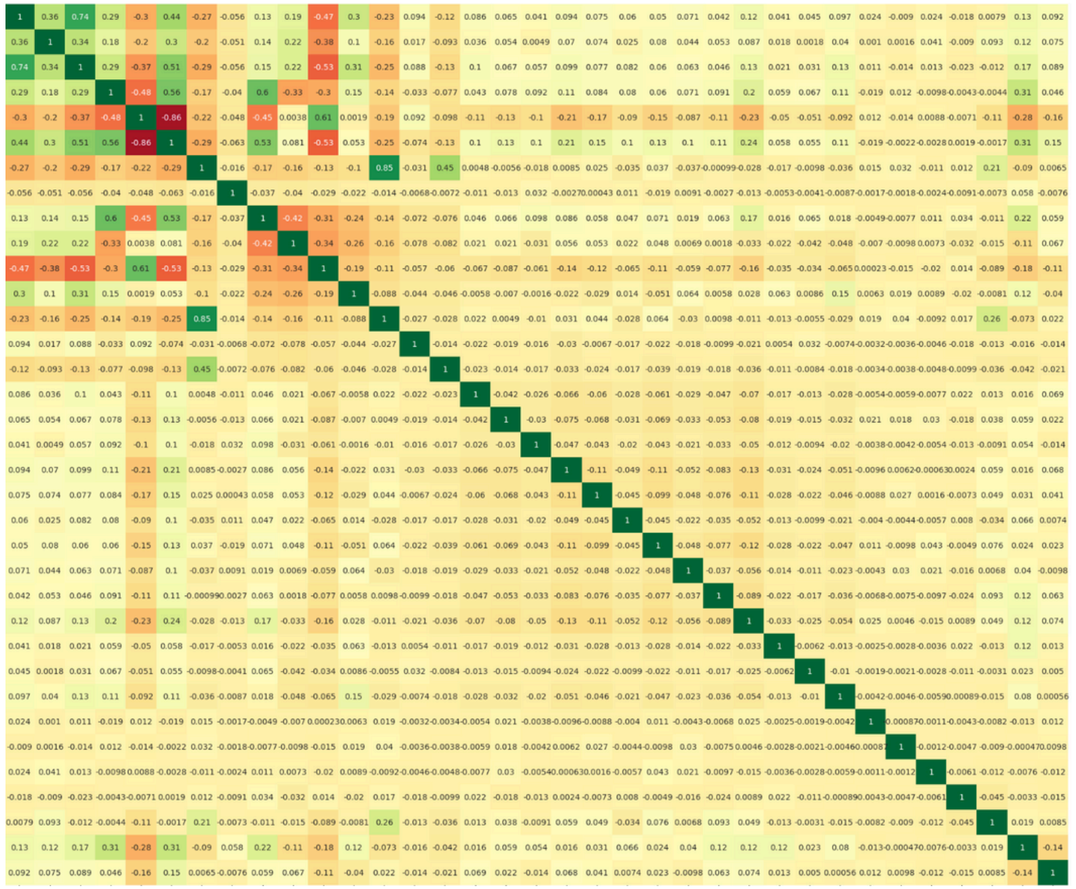
One hot encoding was used to create dummies for the categorical variables. The original categorical variables were dropped along with baseline layers that were a majority or were collinear.

Data was split to train and test in the 70:30 ratio. 6314 records for the train set and 2706 records for the test set.

Standard Scaler from Sklearn was used to scale the numerical variables as this is the better option for logistic regression. (normally distributed variables)

Variables were not dropped based on correlation matrix as we were going to use RFE to select top ‘n’ variables.

Correlation Heatmap of the final variables for modelling



Modelling

Dep. Variable:	Converted	No. Observations:	6314
Model:	GLM	Df Residuals:	6279
Model Family:	Gaussian	Df Model:	34
Link Function:	Identity	Scale:	0.15294
Method:	IRLS	Log-Likelihood:	-3013.6
Date:	Tue, 18 Feb 2025	Deviance:	960.29
Time:	22:07:08	Pearson chi2:	960.
No. Iterations:	3	Pseudo R-squ. (CS):	0.4200
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.8513	0.123	6.896	0.000	0.609	1.093
TotalVisits	0.0361	0.008	4.724	0.000	0.021	0.051
Website_time	0.2074	0.006	36.598	0.000	0.196	0.219
Views_Per_Visit	-0.0255	0.008	-3.109	0.002	-0.042	-0.009
Free_E_Book_Chosen	-0.0370	0.016	-2.352	0.019	-0.068	-0.006
Lead_Origin_API	-0.7098	0.082	-8.700	0.000	-0.870	-0.550
Lead_Origin_Landing Page Submission	-0.8091	0.082	-9.816	0.000	-0.971	-0.648
Lead_Origin_Lead Import	-0.4997	0.149	-3.356	0.001	-0.792	-0.208
Lead_Source_Direct Traffic	0.0606	0.103	0.591	0.554	-0.140	0.262
Lead_Source_Google	0.1099	0.102	1.077	0.282	-0.090	0.310
Lead_Source_Olark Chat	0.2409	0.103	2.332	0.020	0.038	0.443
Lead_Source_Organic Search	0.1100	0.103	1.069	0.285	-0.092	0.312
Lead_Source_Reference	-0.0232	0.125	-0.185	0.853	-0.269	0.222
Lead_Source_Referral Sites	0.0597	0.111	0.538	0.590	-0.158	0.277
Lead_Source_Welingak Website	0.2852	0.130	2.201	0.028	0.031	0.539
Specialization_Banking, Investment And Insurance	0.1976	0.031	6.287	0.000	0.136	0.259
Specialization_Business Administration	0.1446	0.029	4.963	0.000	0.088	0.202
Specialization_E-Business	0.1532	0.041	3.744	0.000	0.073	0.233
Specialization_Finance	0.1479	0.023	6.332	0.000	0.102	0.194
Specialization_Human Resource	0.1431	0.024	6.038	0.000	0.097	0.190
Specialization_International Business	0.1504	0.039	3.835	0.000	0.074	0.227
Specialization_Marketing	0.1666	0.023	7.137	0.000	0.121	0.212
Specialization_Media and Advertising	0.1875	0.037	5.026	0.000	0.114	0.261
Specialization_Operations	0.1643	0.027	6.049	0.000	0.111	0.218
Specialization_Other Management	0.1401	0.023	6.052	0.000	0.095	0.185
Specialization_Rural and Agribusiness	0.1165	0.059	1.984	0.047	0.001	0.232
Specialization_Services Excellence	0.0651	0.074	0.880	0.379	-0.080	0.210
Specialization_Travel and Tourism	0.1566	0.039	4.033	0.000	0.080	0.233
Current_Occupation_Businessman	-0.1120	0.175	-0.638	0.523	-0.456	0.232
Current_Occupation_Housewife	0.4612	0.160	2.879	0.004	0.147	0.775
Current_Occupation_Other	-0.1470	0.124	-1.182	0.237	-0.391	0.097
Current_Occupation_Student	0.0458	0.033	1.392	0.164	-0.019	0.110
Current_Occupation_Working Professional	0.3689	0.020	18.755	0.000	0.330	0.407
City_Other Cities	0.0438	0.015	2.990	0.003	0.015	0.072
City_Thane & Outskirts	0.0100	0.019	0.526	0.599	-0.027	0.047

This is the first model built with the final set of variables chosen after data preparation.

34 variables were used to construct the model using ‘statsmodels’ library.

We can see that multiple variables are insignificant and did not add to the performance of the model.

So, in the next step RFE was used to select the top 15 variables best representing the data.

```
Accuracy Score: 0.79
Sensitivity: 0.63
Specificity: 0.89
FPR: 0.11
Precision: 0.77
```

- 15 variables were selected using RFE and the below was the model that resulted.
- All the VIF values are well below the standard limit of 5 which means there is very less collinearity in the model.
- For a threshold of 0.5, the model has decent accuracy and precision but is not capturing enough conversions.
- Also, some variables are still insignificant.

Features	VIF
Lead_Source_Olark Chat	2.26
Lead_Origin_API	2.08
Lead_Origin_Landing Page Submission	1.77
Website_time	1.28
Specialization_Finance	1.27
Specialization_Marketing	1.22
Lead_Source_Reference	1.21
Current_Occupation_Working Professional	1.21
Specialization_Operations	1.14
Specialization_Banking, Investment And Insurance	1.09
Specialization_Media and Advertising	1.06
Specialization_Travel and Tourism	1.06
Lead_Source_Welingak Website	1.01
Lead_Origin_Lead Import	1.00
Current_Occupation_Housewife	1.00

```
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6314
Model:                            GLM   Df Residuals:                     6298
Model Family:                Binomial   Df Model:                           15
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2963.7
Date:                Tue, 18 Feb 2025   Deviance:                       5927.3
Time:                        22:22:55   Pearson chi2:                 6.62e+03
No. Iterations:                    21   Pseudo R-squ. (CS):             0.3220
Covariance Type:            nonrobust
==============================================================================================
                                                 coef    std err          z      P>|z|
----------------------------------------------------------------------------------------------
const                                          2.7286      0.625      4.364      0.000
Website_time                                   1.1300      0.038     29.858      0.000
Lead_Origin_API                               -3.9001      0.630     -6.189      0.000
Lead_Origin_Landing Page Submission           -4.1501      0.629     -6.598      0.000
Lead_Origin_Lead Import                       -2.9542      0.800     -3.692      0.000
Lead_Source_Olark Chat                         0.7829      0.110      7.122      0.000
Lead_Source_Reference                          0.2645      0.668      0.396      0.692
Lead_Source_Welingak Website                   2.7020      1.184      2.282      0.022
Specialization_Banking, Investment And Insurance  0.6262      0.173      3.627      0.000
Specialization_Finance                         0.3015      0.114      2.653      0.008
Specialization_Marketing                       0.4028      0.117      3.441      0.001
Specialization_Media and Advertising           0.5152      0.212      2.433      0.015
Specialization_Operations                      0.4422      0.147      2.999      0.003
Specialization_Travel and Tourism              0.3450      0.217      1.590      0.112
Current_Occupation_Housewife                  23.1263   1.62e+04      0.001      0.999
Current_Occupation_Working Professional        2.8922      0.181     15.990      0.000
==============================================================================================
```

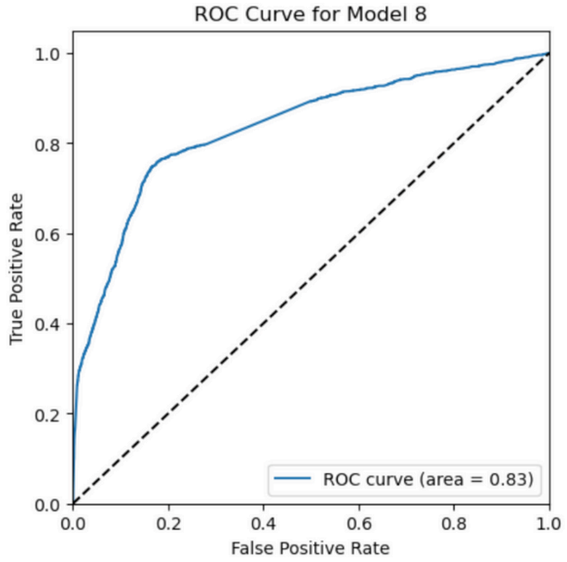
Accuracy Score: 0.79
Sensitivity: 0.63
Specificity: 0.88
FPR: 0.12
Precision: 0.77

- After 6 iterations of removing insignificant variables, we have arrived at a model with significant and non-collinear variables with decent metrics.
- By optimising the threshold, we can improve the metrics of the model.

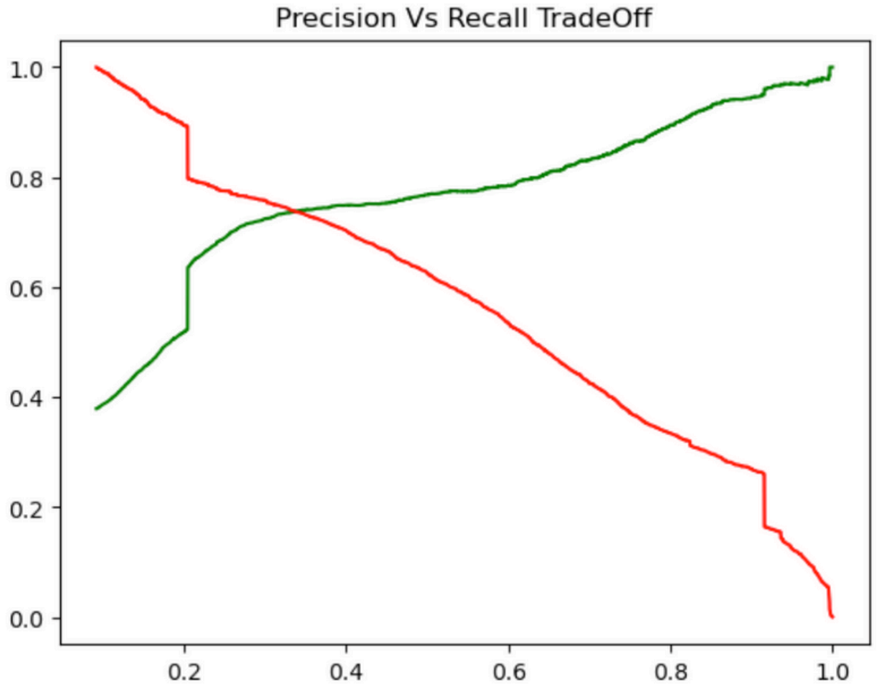
Dep. Variable:	Converted	No. Observations:	6314
Model:	GLM	Df Residuals:	6304
Model Family:	Binomial	Df Model:	9
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2981.1
Date:	Tue, 18 Feb 2025	Deviance:	5962.1
Time:	22:38:09	Pearson chi2:	6.82e+03
No. Iterations:	6	Pseudo R-squ. (CS):	0.3182
Covariance Type:	nonrobust		

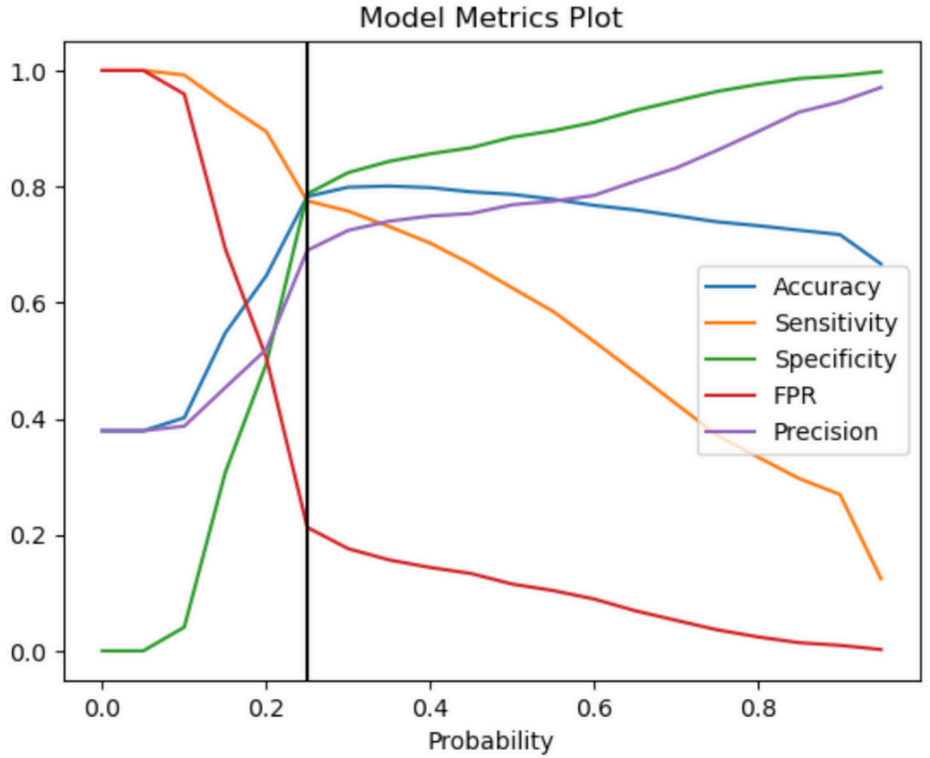
	coef	std err	z	P>|z|
const	3.3703	0.215	15.648	0.000
Website_time	1.1289	0.038	29.935	0.000
Lead_Origin_API	-4.5167	0.229	-19.689	0.000
Lead_Origin_Landing Page Submission	-4.6773	0.223	-20.954	0.000
Lead_Origin_Lead Import	-3.5362	0.543	-6.508	0.000
Lead_Source_Olark Chat	0.7718	0.110	7.045	0.000
Specialization_Banking, Investment And Insurance	0.5211	0.171	3.056	0.002
Specialization_Marketing	0.2974	0.114	2.607	0.009
Specialization_Operations	0.3417	0.145	2.359	0.018
Current_Occupation_Working Professional	2.9039	0.181	16.055	0.000

Features	VIF
Lead_Source_Olark Chat	2.25
Lead_Origin_API	2.06
Lead_Origin_Landing Page Submission	1.32
Website_time	1.21
Specialization_Marketing	1.14
Specialization_Operations	1.10
Current_Occupation_Working Professional	1.08
Specialization_Banking, Investment And Insurance	1.06
Lead_Origin_Lead Import	1.00



- Our final model captures 83% of the variance in the data.
- A plot of the metrics for all possible thresholds (between 0 and 1) intersects at 0.25 which is our optimal cut-off based on the plot.
- However, since we're trying to optimize for the 'recall' value to be higher, a cut-off between 0.25 and 0.3 was tested for.



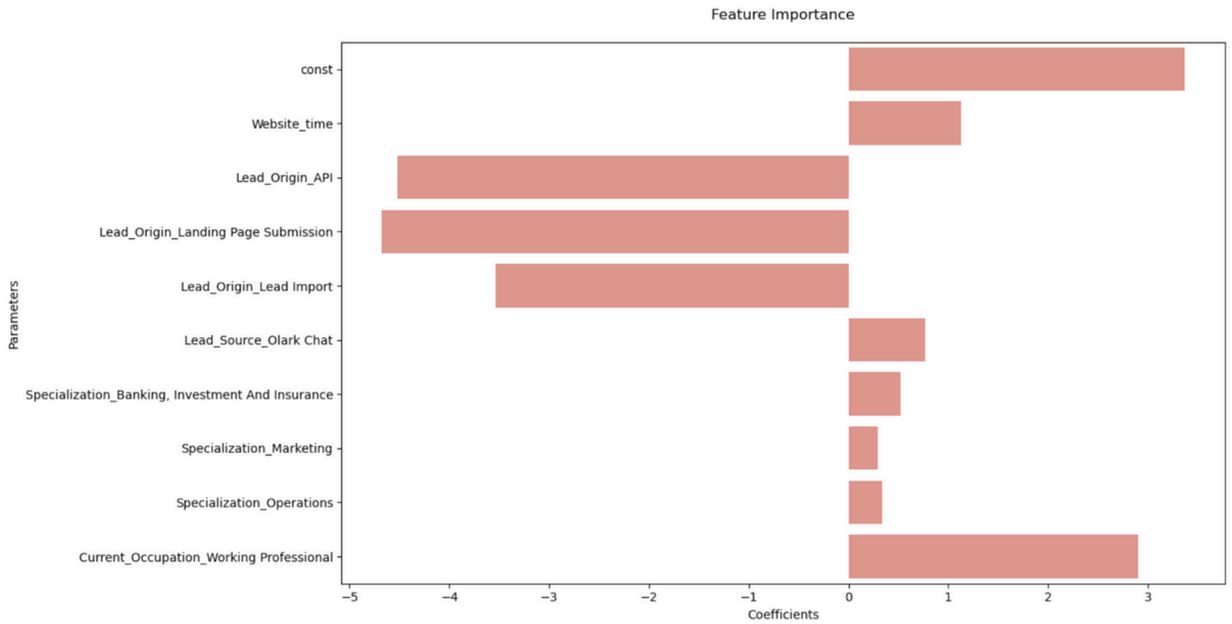


Test Predictions and Feature Importance

```
Accuracy Score: 0.79
Sensitivity: 0.75
Specificity: 0.81
FPR: 0.19
Precision: 0.71
```

- After testing between 0.25 and 0.3, 0.28 gave us the best recall value while still maintaining good accuracy and precision.
- False Positives are also quite low which will lower the wastage of resources for the company.
- The final model has 9 predicting variables.

Parameters	Coefficients
Current_Occupation_Working Professional	2.903854
Website_time	1.128875
Lead_Source_Olark Chat	0.771808
Specialization_Banking, Investment And Insurance	0.521116
Specialization_Operations	0.341743
Specialization_Marketing	0.297422
Lead_Origin_Lead Import	-3.536228
Lead_Origin_API	-4.516706
Lead_Origin_Landing Page Submission	-4.677257



Lead Score

The formula to score the leads based on our model out of 100 is:

$$\text{Lead Score (out of 100)} = \frac{100}{1 + e^{-\left(\text{Intercept} + \sum_{i=1}^{n} \text{Coefficient}_i \cdot \text{Feature}_i\right)}}$$

```python
#Python Code for a Lead Score
params = leads_mod8.fit().params
intercept = params['const']
coefficients = params.drop('const')

def lead_score(row):
    log = intercept
    for i,j in coefficients.items():
        log = log + row[i]*j
    prob = 1/(1+np.exp(-log))

    return round(prob*100)
```

Suggestions

1. Optimize Sales Team Allocation
Focus on high-scoring leads (e.g., leads with scores above 75) to ensure the sales team spends their time and effort on the most promising prospects, improving conversion efficiency.

2. Refine Marketing Efforts
Redirect marketing budgets toward high-performing channels, such as Google Ads, direct traffic, and organic search, which generated the majority of the successful leads.

3. Enhance Olark Chat Conversion
Revamp the chat experience by training representatives on product knowledge, improving response times, and integrating AI-based chatbots for better lead engagement.

4. Value-Add to Free E-Book
Improve the content and value of the free e-book by including career success stories, industry-relevant trends, and actionable insights to build trust and credibility.

5. Target Career Aspirants
Develop campaigns tailored to unemployed professionals and those seeking career advancement, emphasizing how the courses address "better career prospects."

6. Invest in Training the Sales Team
Equip the sales team with training on data-driven lead prioritization and strategies to handle specific personas, such as housewives or non-specialized users.

7. Personalize Communication
Use score segmentation and other lead characteristics above to personalize communication for leads, improving engagement and trust in the company's offerings.