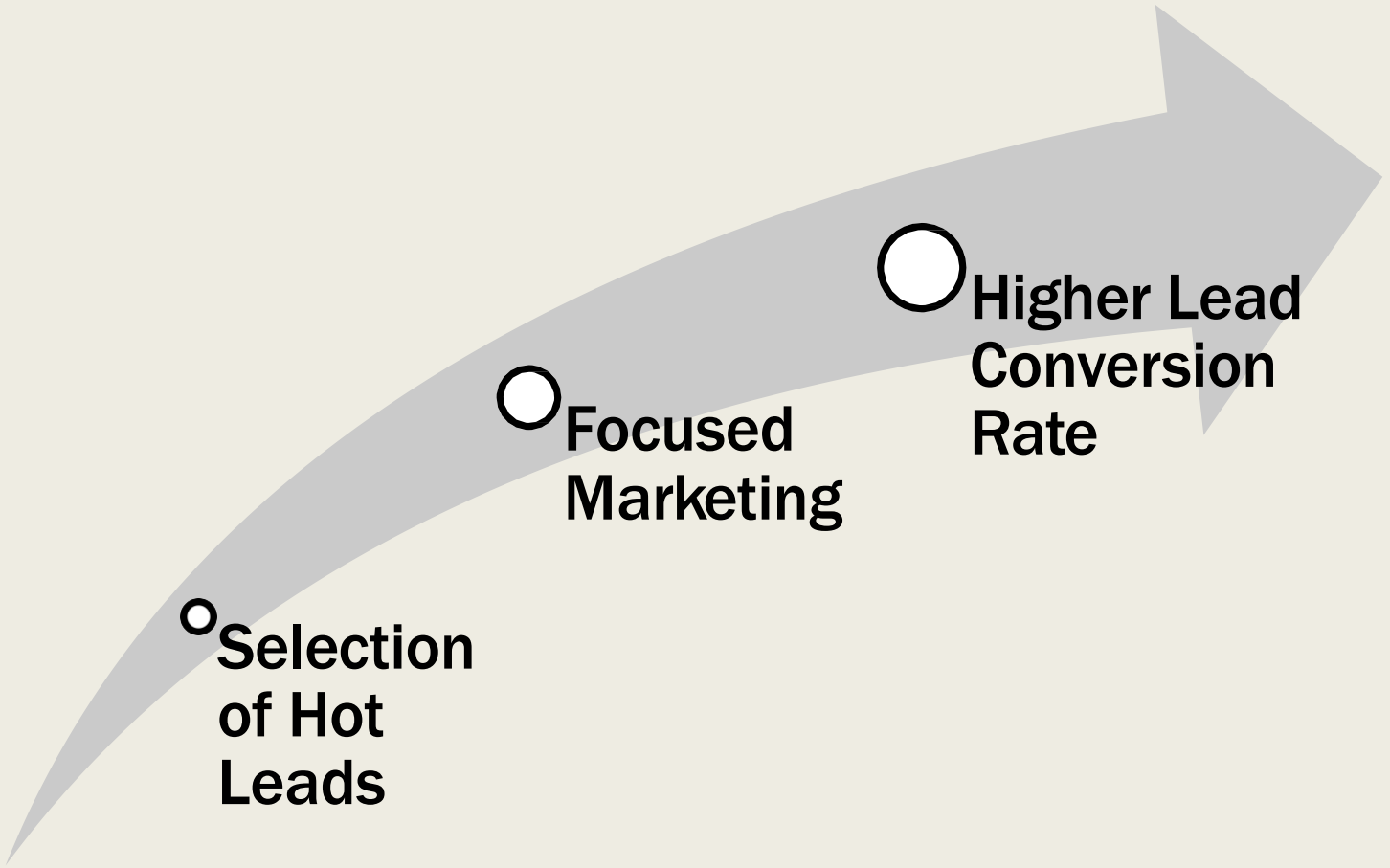
LEAD SCORING CASE STUDY

Focused business approach using logistic regression technique

Akash R
Swaroop
Sundeep Allu

Business Objective

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



METHODOLOGY

To build a Logistic Regression model that assigns lead scores to all leads such that the customers with higher lead score have a higher conversion chance and vice versa.
Target Lead Conversion Rate ≈ 80%

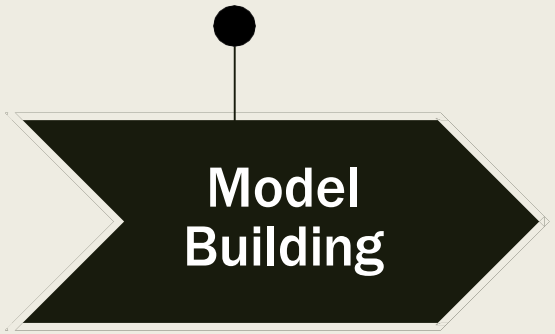
Importing and Observing the past data provided by the Company

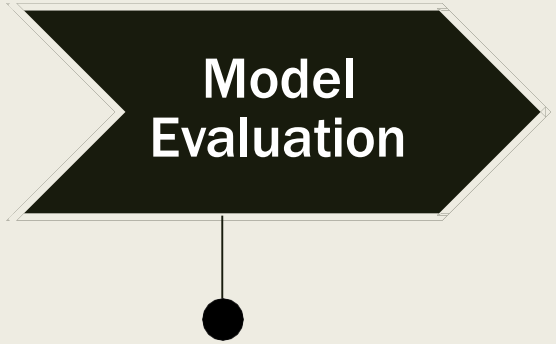
Univariate and Bivariate analysis

- Missing value imputation
- Removing duplicate data and other redundancies

- Outlier treatment
- Dropping unnecessary columns
- Dummy variable creation
- Feature standardization

- Feature selection using RFE
- Manual feature elimination based on p-values and VIFs





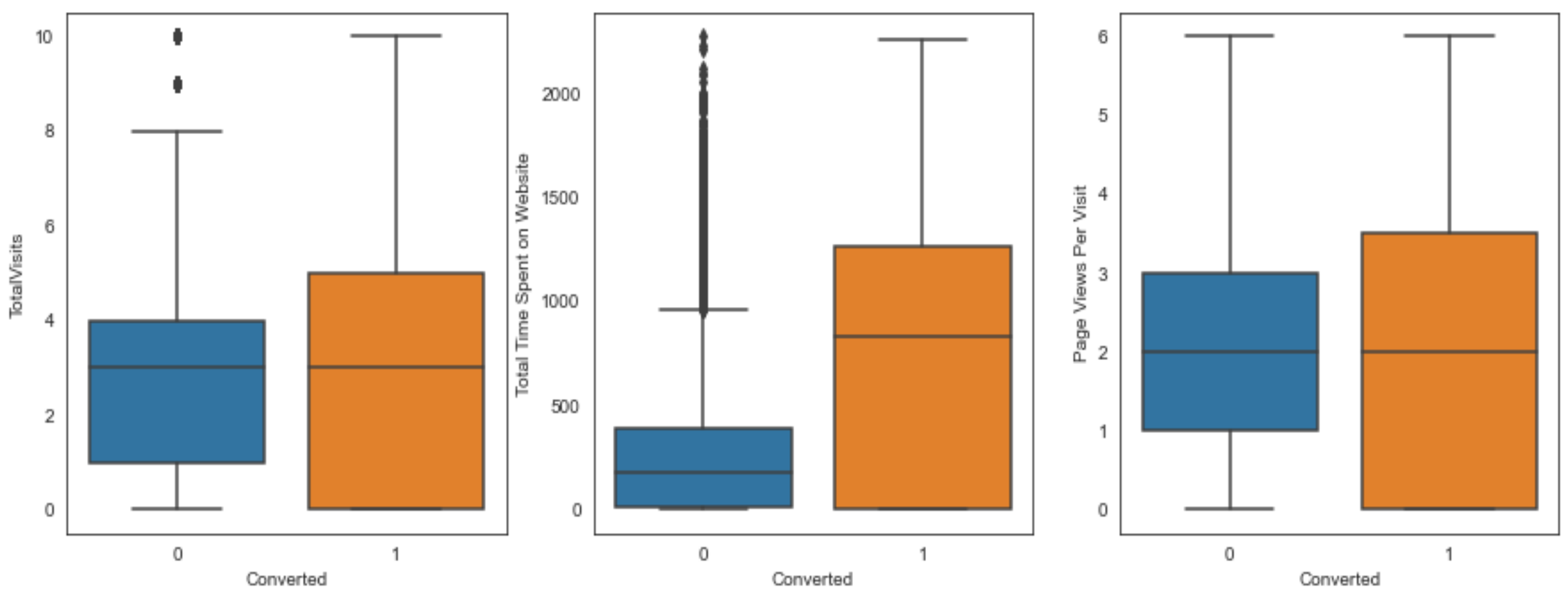
- Evaluating model based on various evaluation metrics
- Finding the optimal probability threshold

- Finalizing the first model
- Using predicted probabilities to calculate Lead Scores:
  Lead Score = Probability* 100

DATA VISUALIZATION

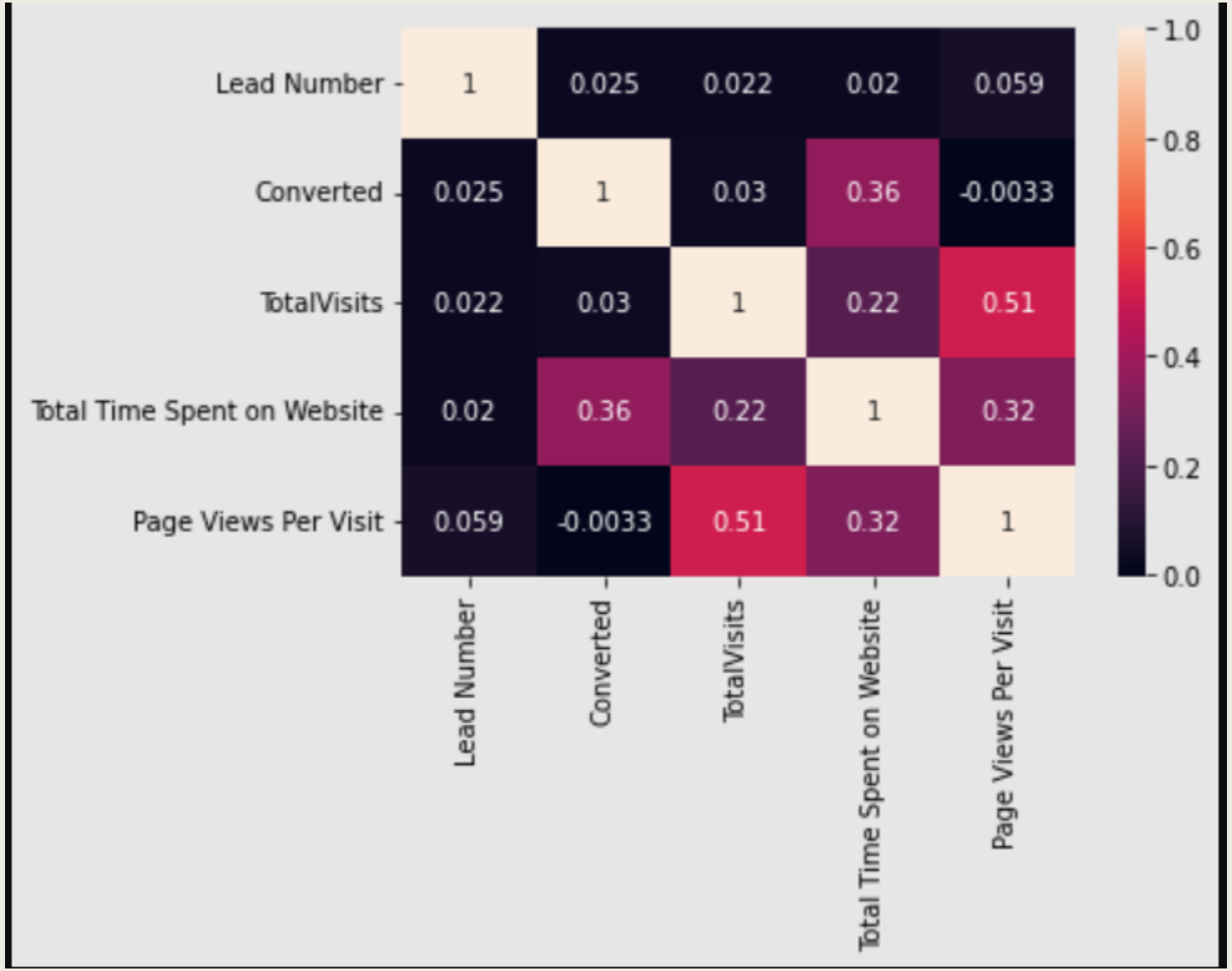
- To identify important features
- To get insights

Numerical Variables

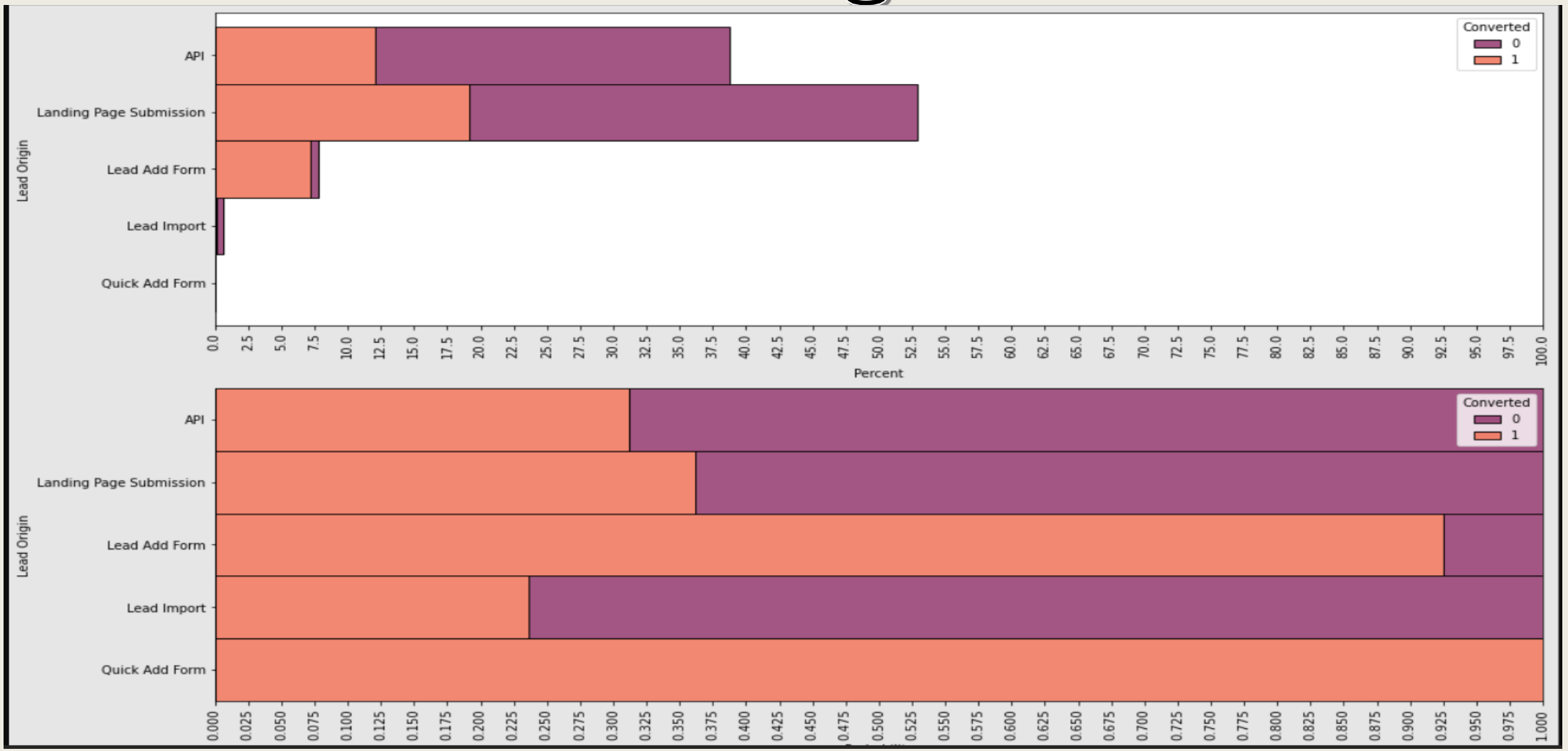


People spending more time on website are more likely to get converted.

HEAT MAP

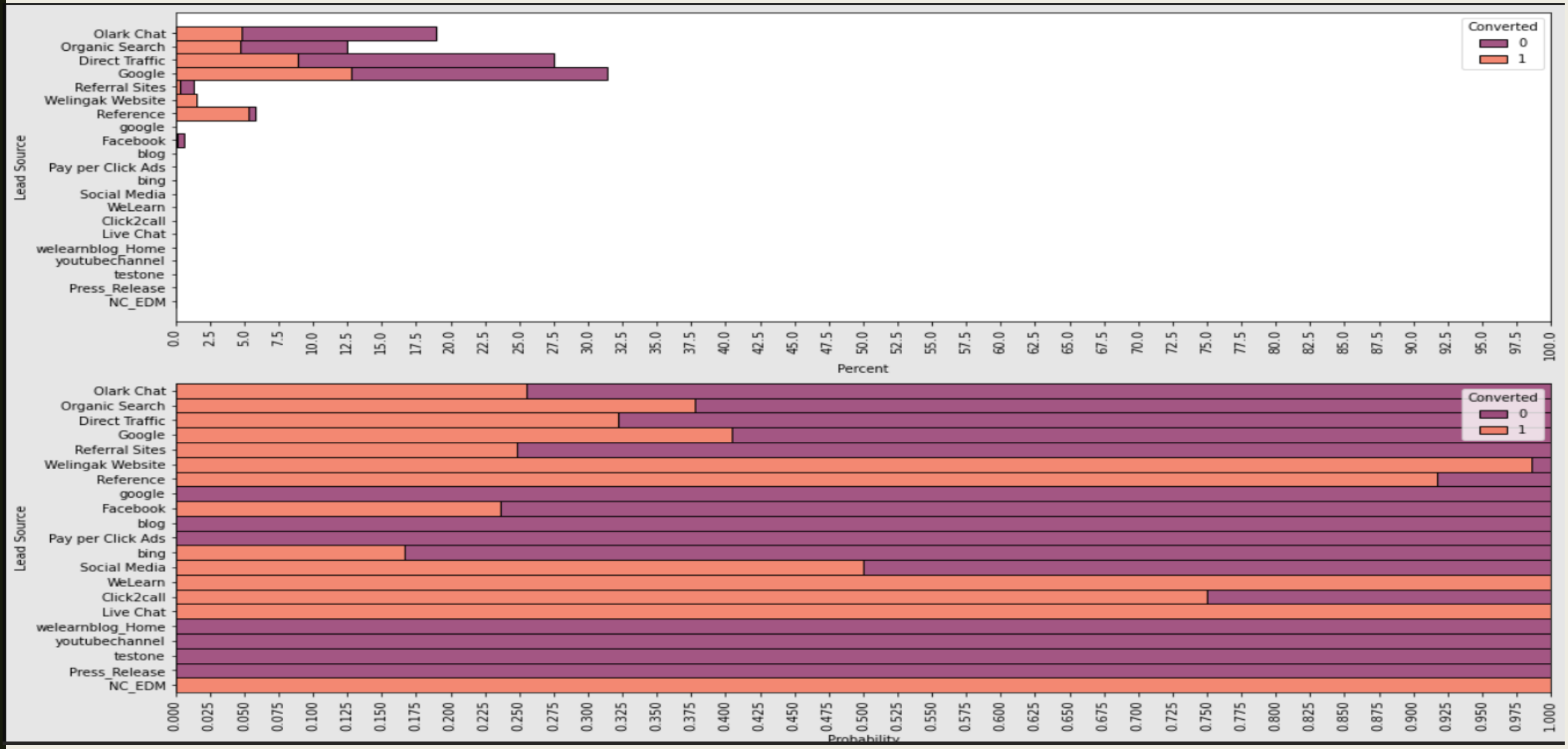

Lead Number
Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
1
0.025
0.022
0.02
0.059
0.025
1
0.03
0.36
-0.0033
0.022
0.03
1
0.22
0.51
0.02
0.36
0.22
1
0.32
0.059
-0.0033
0.51
0.32
1
Lead Number
Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
1.0
0.8
0.6
0.4
0.2
0.0

Lead Origin



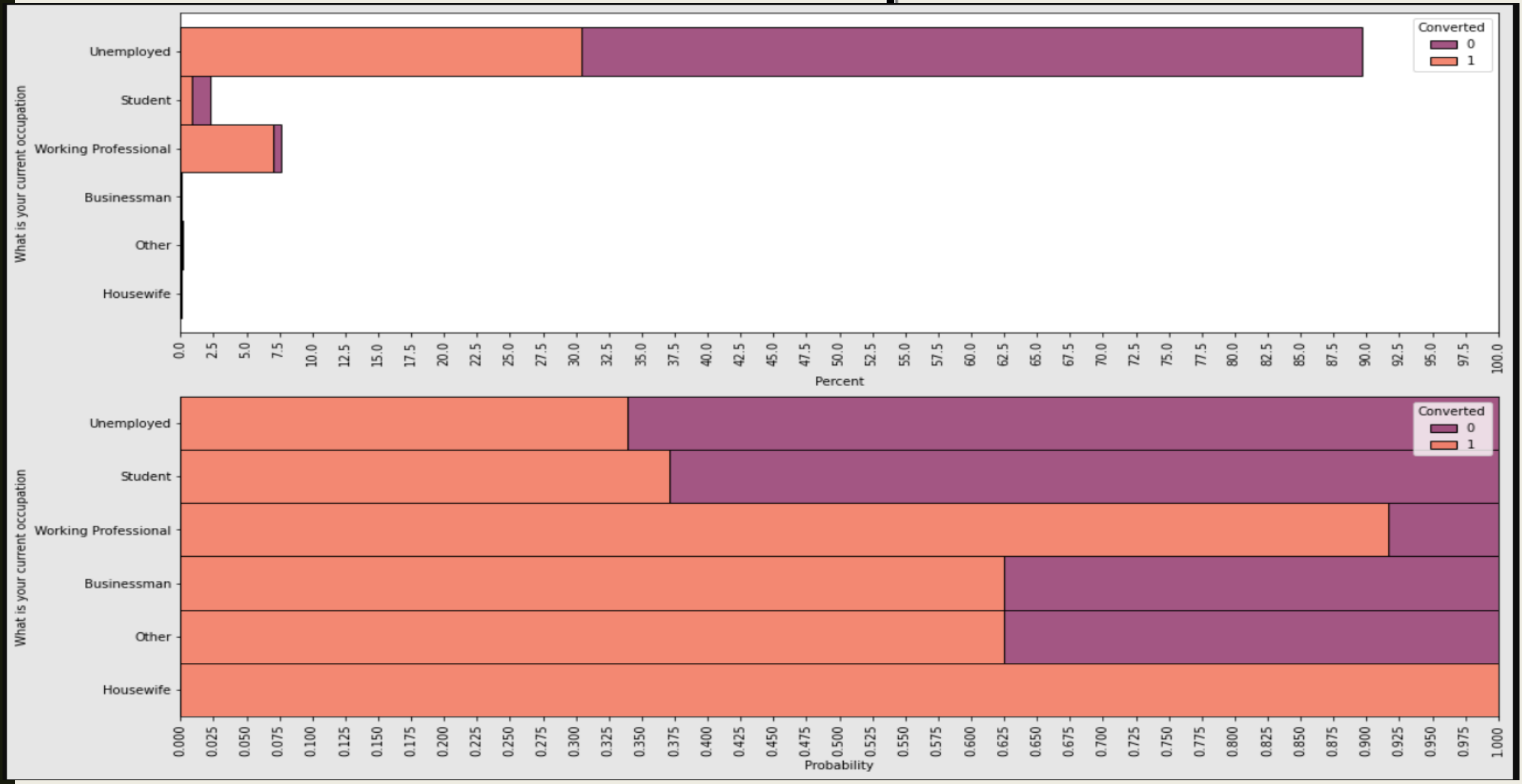
- ***'API'*** and '***Landing Page Submission***' generate the most leads but have less conversion rates, whereas '***Lead Add Form***' generates less leads but conversion rate is great.
- **Try to increase conversion rate for 'API' and 'Landing Page Submission', and increase leads generation using 'Lead Add Form'.**

Lead Source



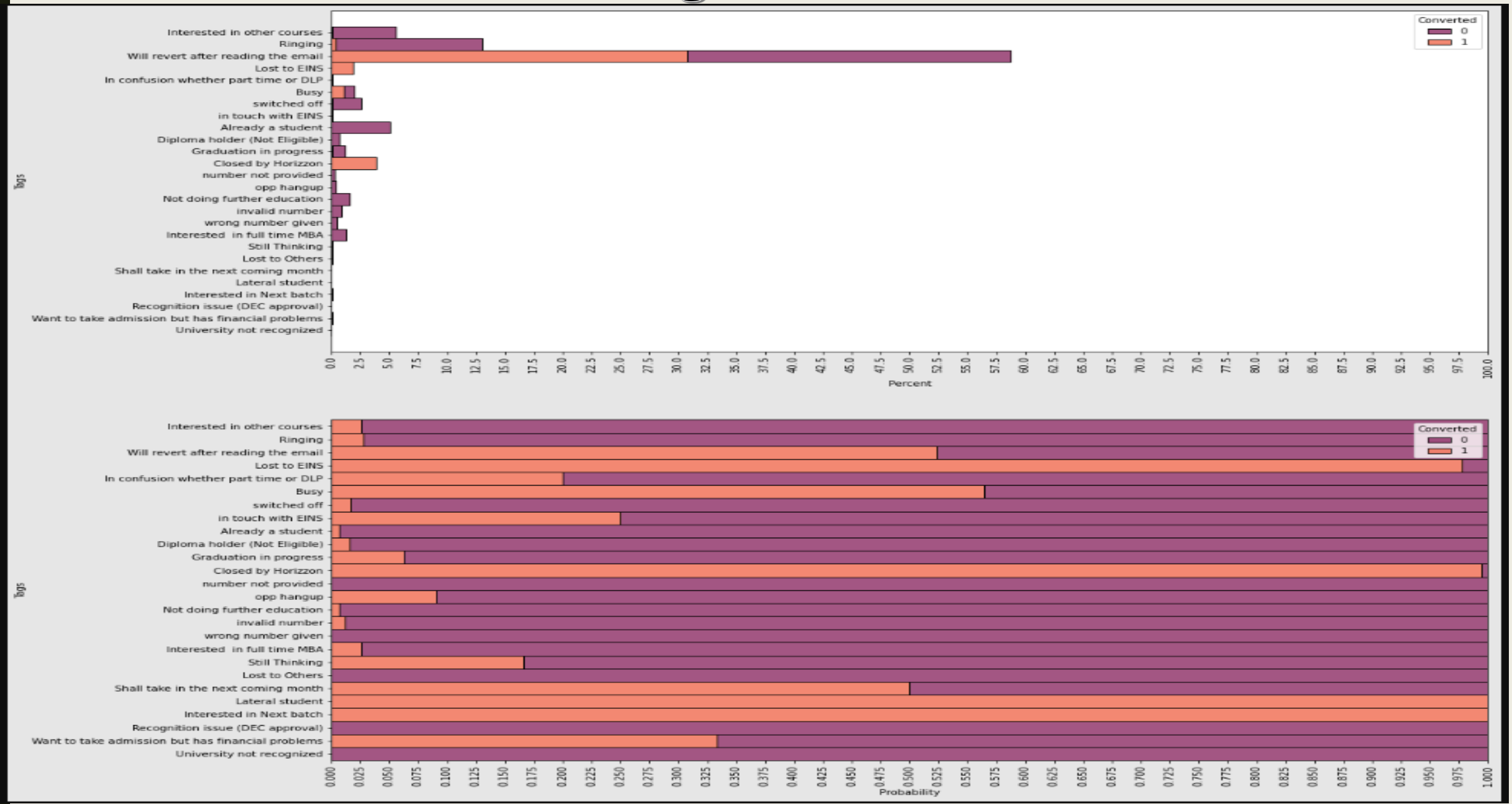
- Very high conversion rates for lead sources '**Reference**' and '**Welingak Website**'.
- Most leads are generated through '**Direct Traffic**' and '**Google**'.

Current Occupation



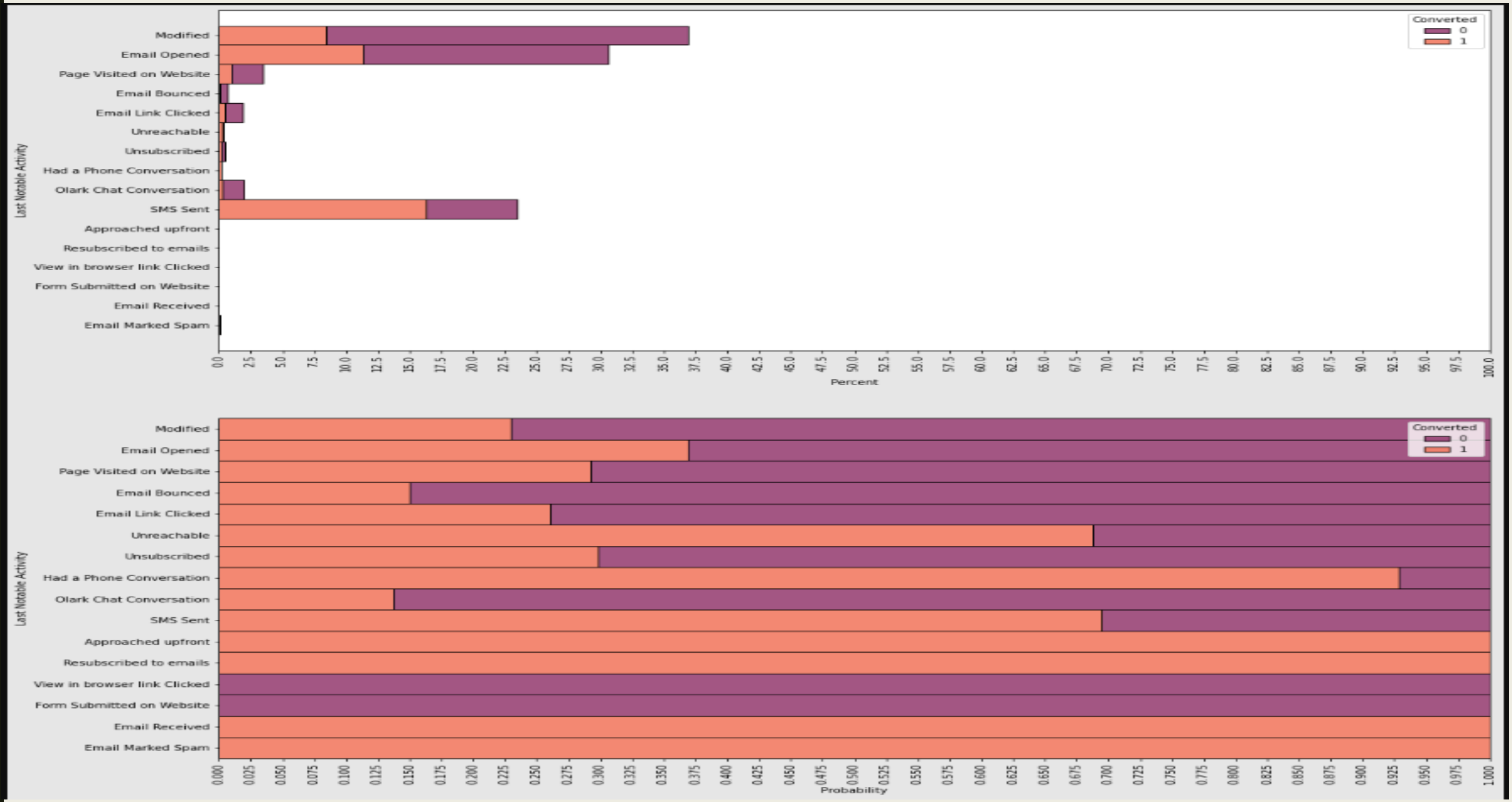
Working Professionals are most likely to get converted.

Tags



High conversion rates for tags **'Will revert after reading the email', 'Closed by Horizon', 'Lost to EINS', 'Busy**

Last Notable Activity



Highest conversion rate is for the last notable activity '**SMS Sent**'.

MODEL EVALUATION

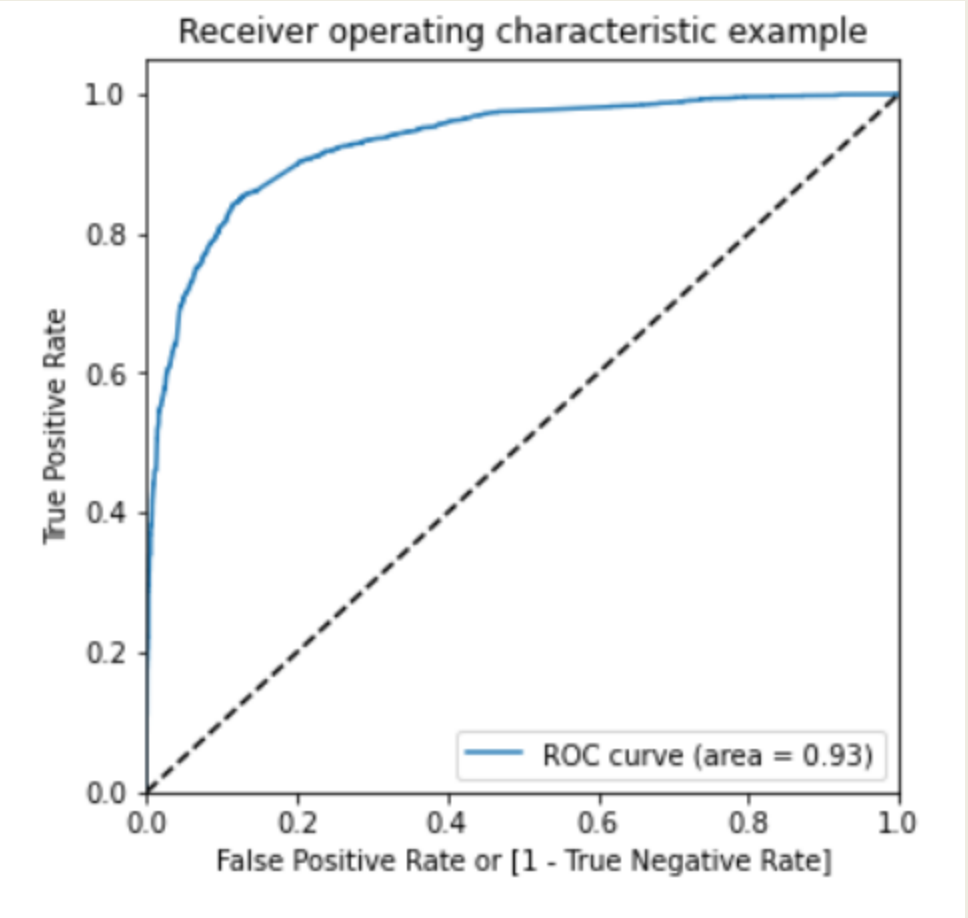
Final Model Summary

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6452
Model Family:	Binomial	Df Model:	15
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2082.9
Date:	Mon, 27 Feb 2023	Deviance:	4165.8
Time:	22:43:30	Pearson chi2:	1.22e+04
No. Iterations:	9	Pseudo R-squ. (CS):	0.4961
Covariance Type:	nonrobust		

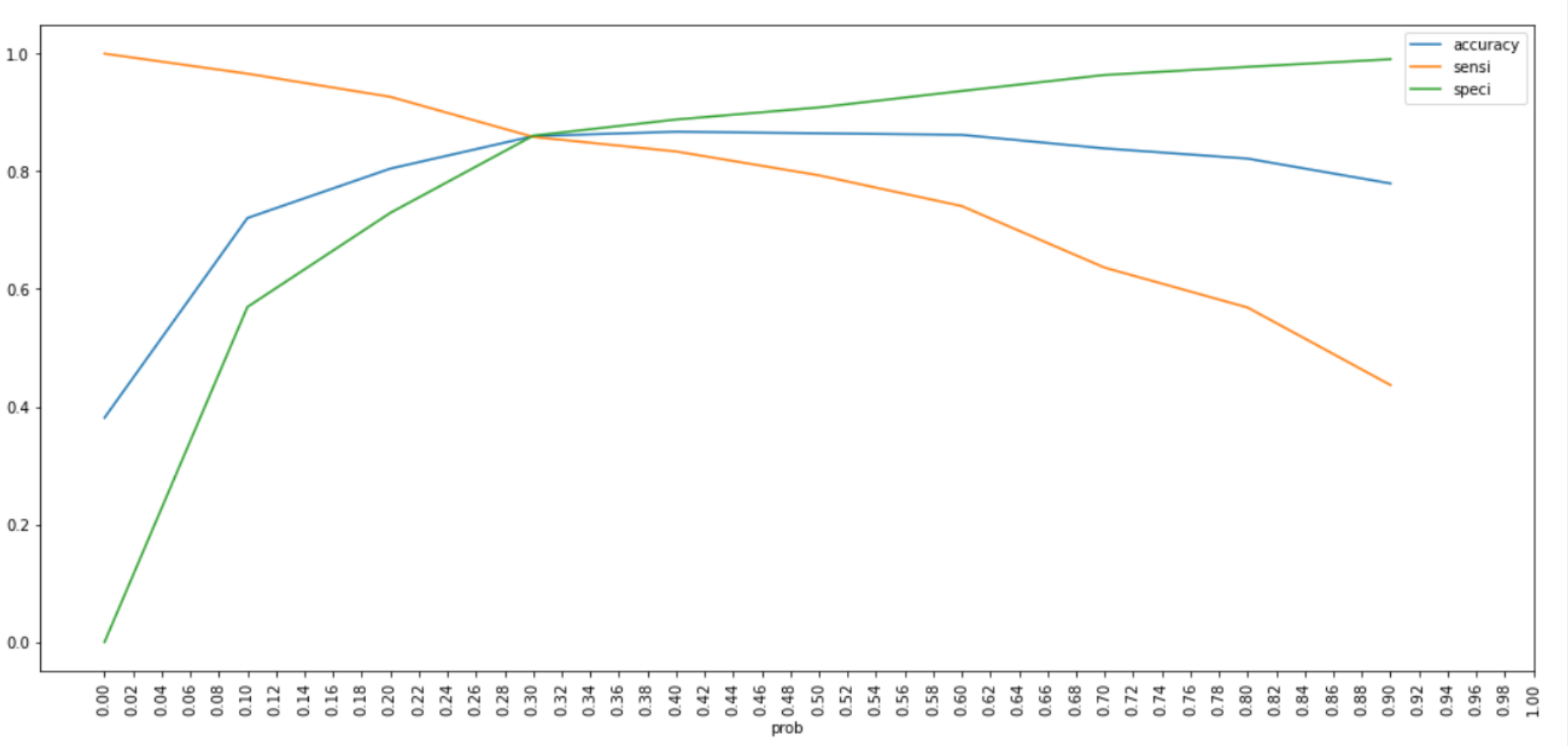
	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.0366	0.093	-21.788	0.000	-2.220	-1.853
TotalVisits	6.2110	2.066	3.006	0.003	2.161	10.261
Total Time Spent on Website	4.6985	0.192	24.453	0.000	4.322	5.075
Lead Origin_Lead Add Form	3.9511	0.220	17.988	0.000	3.521	4.382
Lead Source_Olark Chat	1.0809	0.117	9.257	0.000	0.852	1.310
Do Not Email_Yes	-1.2433	0.179	-6.960	0.000	-1.593	-0.893
Last Activity_SMS Sent	1.7409	0.087	19.917	0.000	1.570	1.912
What is your current occupation_Working Professional	2.7207	0.210	12.986	0.000	2.310	3.131
Tags_Closed by Horizzon	6.2084	1.019	6.093	0.000	4.211	8.205
Tags_Lost to EINS	5.5882	0.730	7.650	0.000	4.157	7.020
Tags_Ringing	-4.1866	0.234	-17.854	0.000	-4.646	-3.727
Last Notable Activity_Email Link Clicked	-1.0197	0.320	-3.190	0.001	-1.646	-0.393
Last Notable Activity_Had a Phone Conversation	3.0772	1.100	2.798	0.005	0.922	5.232
Last Notable Activity_Modified	-1.7345	0.093	-18.746	0.000	-1.916	-1.553
Last Notable Activity_Olark Chat Conversation	-1.7271	0.356	-4.850	0.000	-2.425	-1.029
Last Notable Activity_Page Visited on Website	-0.9929	0.233	-4.257	0.000	-1.450	-0.536

ROC curve



Area under curve = 0.93

Finding Optimal Threshold



Graph showing changes in Sensitivity, Specificity and Accuracy with changes in the probability threshold values

Optimal cutoff = 0.30

Final Results

Data	Train set	Test set
Accuracy	0.859	0.857
Sensitivity	0.858	0.862
Specificity	0.860	0.855
False Positive Rate	0.139	0.145
Positive Predictive Value	0.791	0.795
Negative Predictive Value	0.907	0.904

INFERENCES

Feature Importance

- Three variables which contribute most towards the probability of a lead conversion in decreasing order of impact are:
    1) Total visits
    2) Tags_Closed by Horizzon
    3) Tags_Lost to EINS
- There are dummy features created from the categorical variable Tags.
- All three **contribute positively** towards the probability of a lead conversion.
- These results indicate that the company should **focus more on the leads with these three tags.**

Recommendations

- By referring to the data visualizations, focus on
  - *Increasing the conversion rates for the categories generating more leads and*
  - *Generating more leads for categories having high conversion rates.*
- Pay attention to the relative importance of the features in the model and their positive or negative impact on the probability of conversion.
- Based on varying business needs, modify the probability threshold value for identifying potential leads.

THANK YOU