Presentation

On

Lead Score Case StudY

BY

DeePansh Jha

And

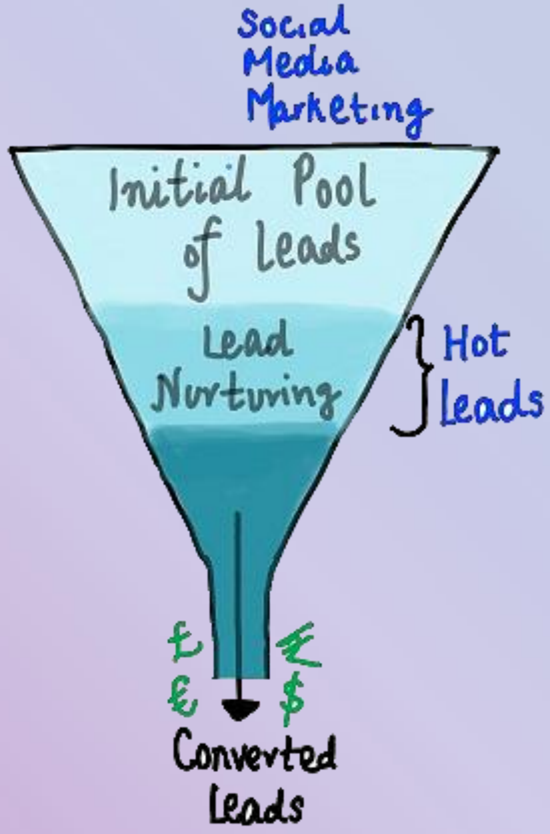
ChaYanika Maiti

Problem Statement

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone. A typical lead conversion process can be represented using the following funnel:



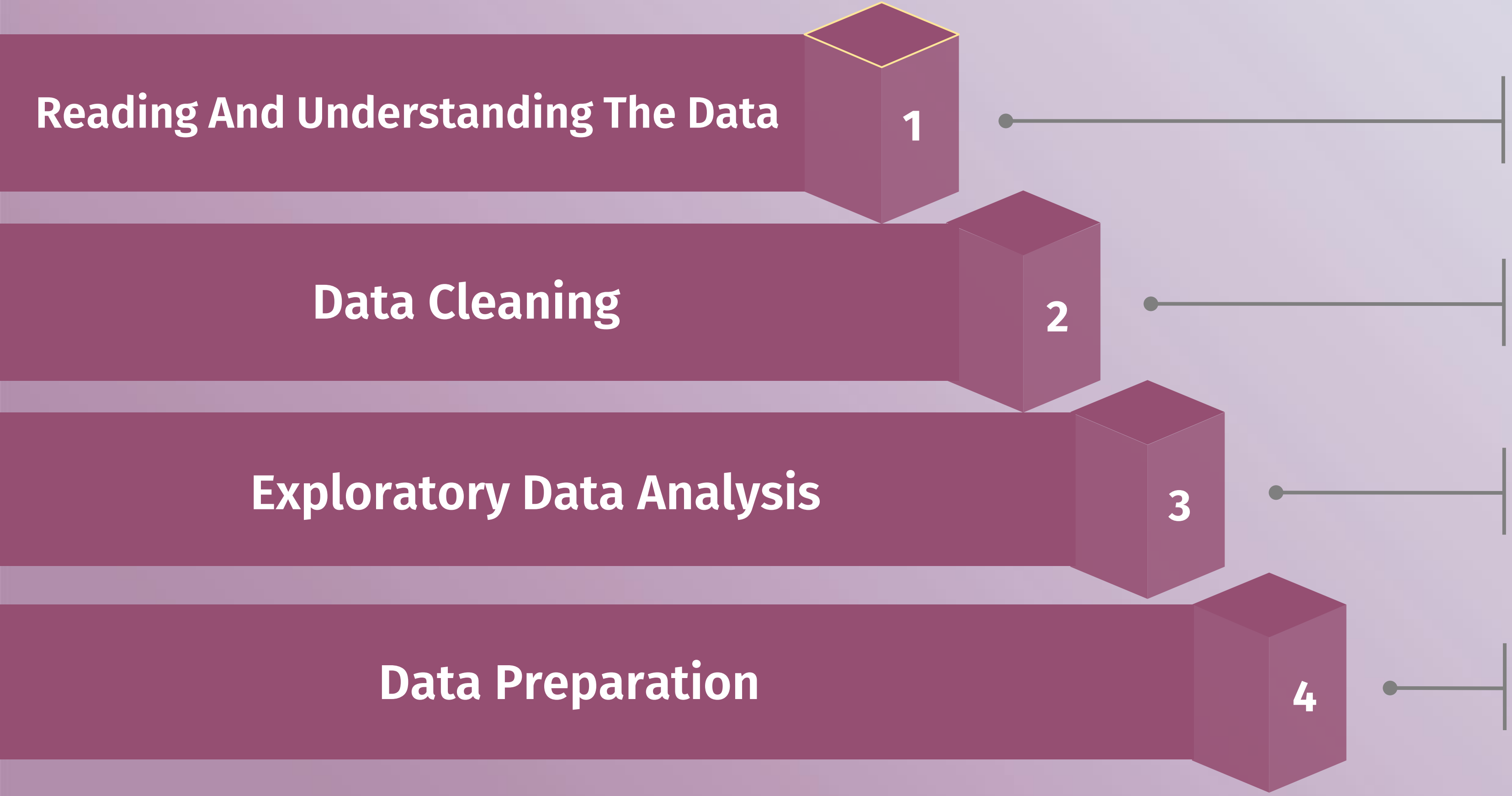
Objectives

- ✓ To help the company **in selecting the** most potential leads, also known as **‘Hot Leads’** whose **Target lead conversion rate is around 80%.**

- ✓ **To build a model wherein a lead score is assigned** to each of the leads such that the customers with **higher lead score** have a **higher conversion** chance and the customers with **lower lead score** have a **lower conversion** chance.

- ✓ **Help** the sales team to divert their focus on **potential leads** & avoid them from making **useless phone calls**.

Methodology Part I



1. Reading And Understanding The Data

Importing and Observing the past data provided by the Company

2. Data Cleaning

- Missing value imputation
- Removing duplicate data and other redundancies

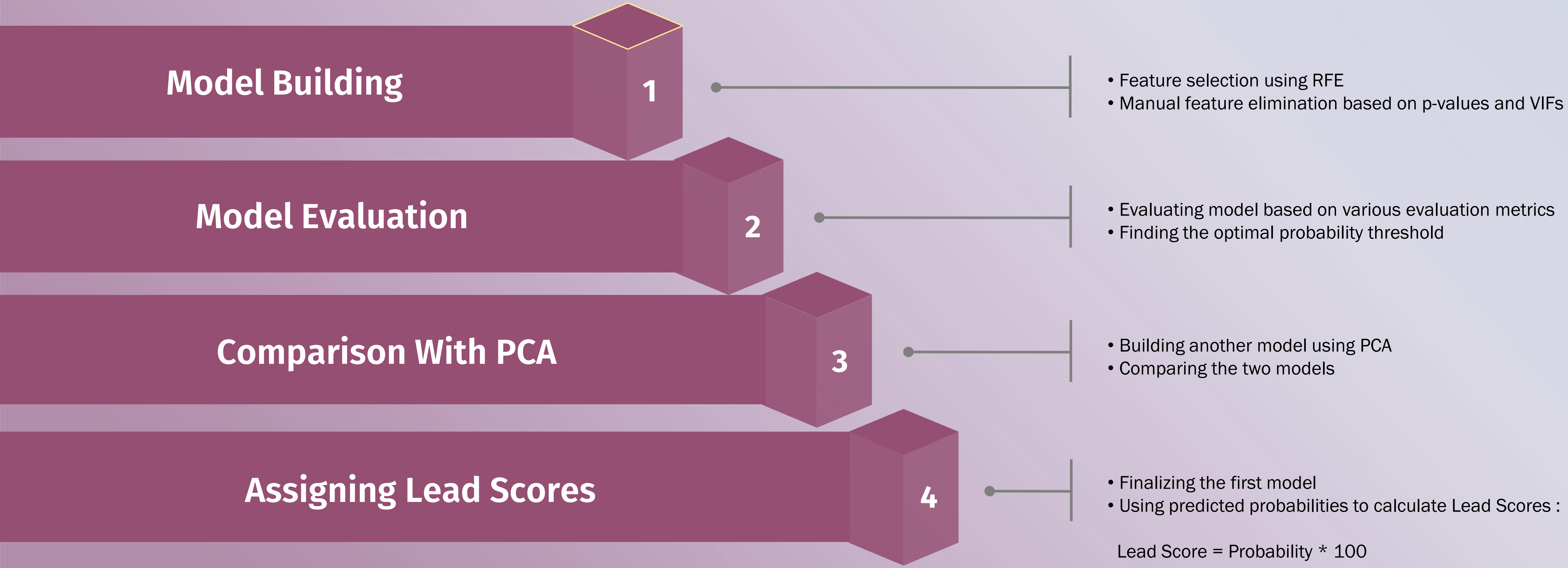
3. Exploratory Data Analysis

Univariate and Bivariate analysis

4. Data Preparation

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

Methodology Part II



Model Building 1

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

Model Evaluation 2

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

Comparison With PCA 3

- Building another model using PCA
- Comparing the two models

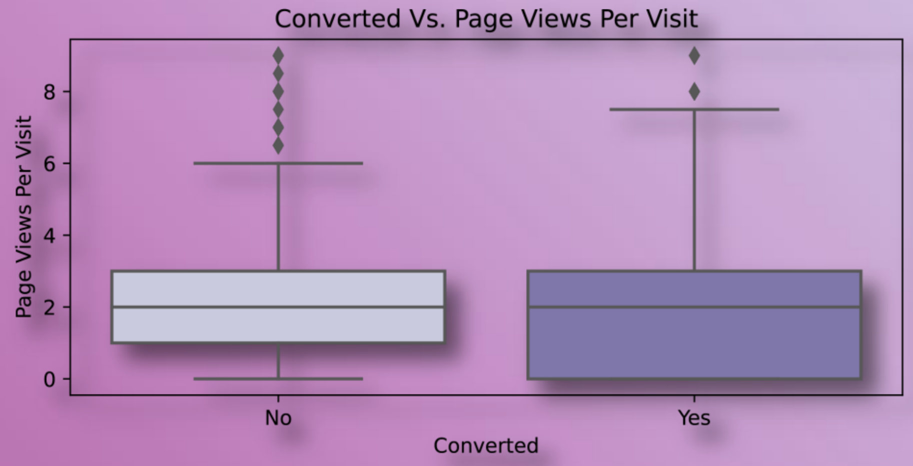
Assigning Lead Scores 4

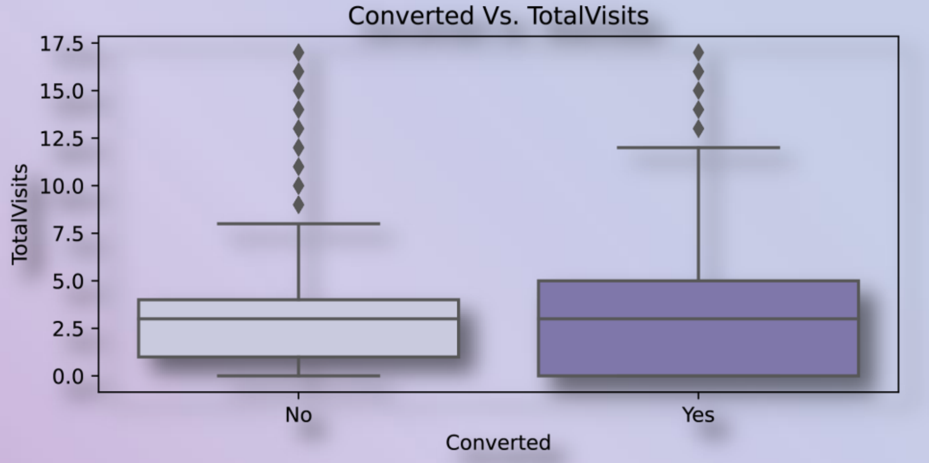
- Finalizing the first model
- Using predicted probabilities to calculate Lead Scores :

Lead Score = Probability * 100

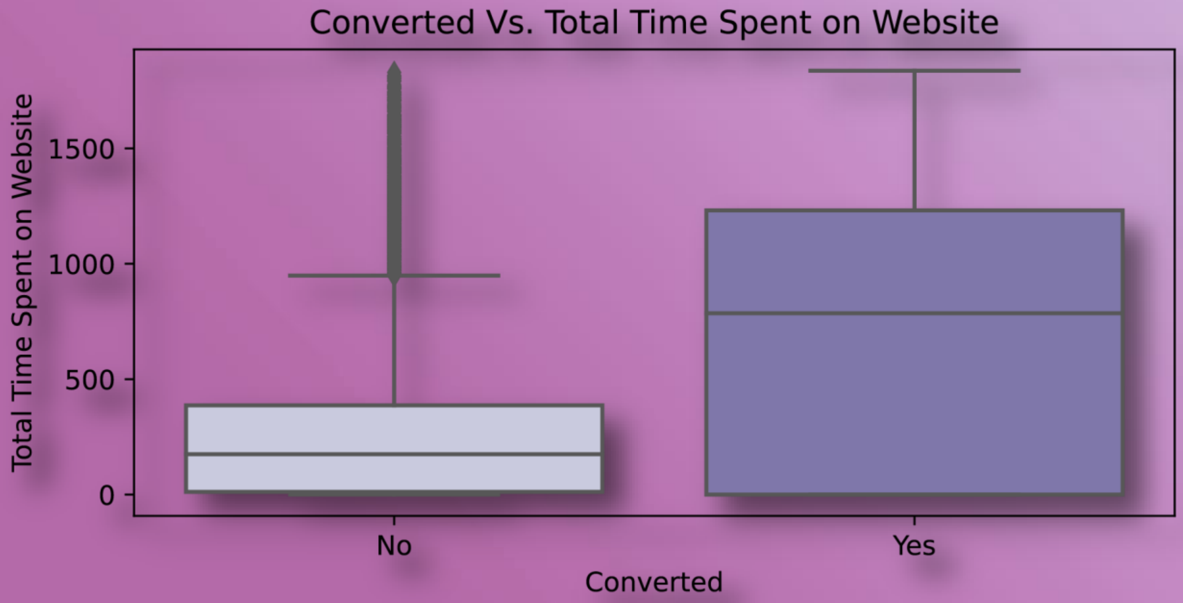
Data Visualization Continuous Variables

Continuous Variables

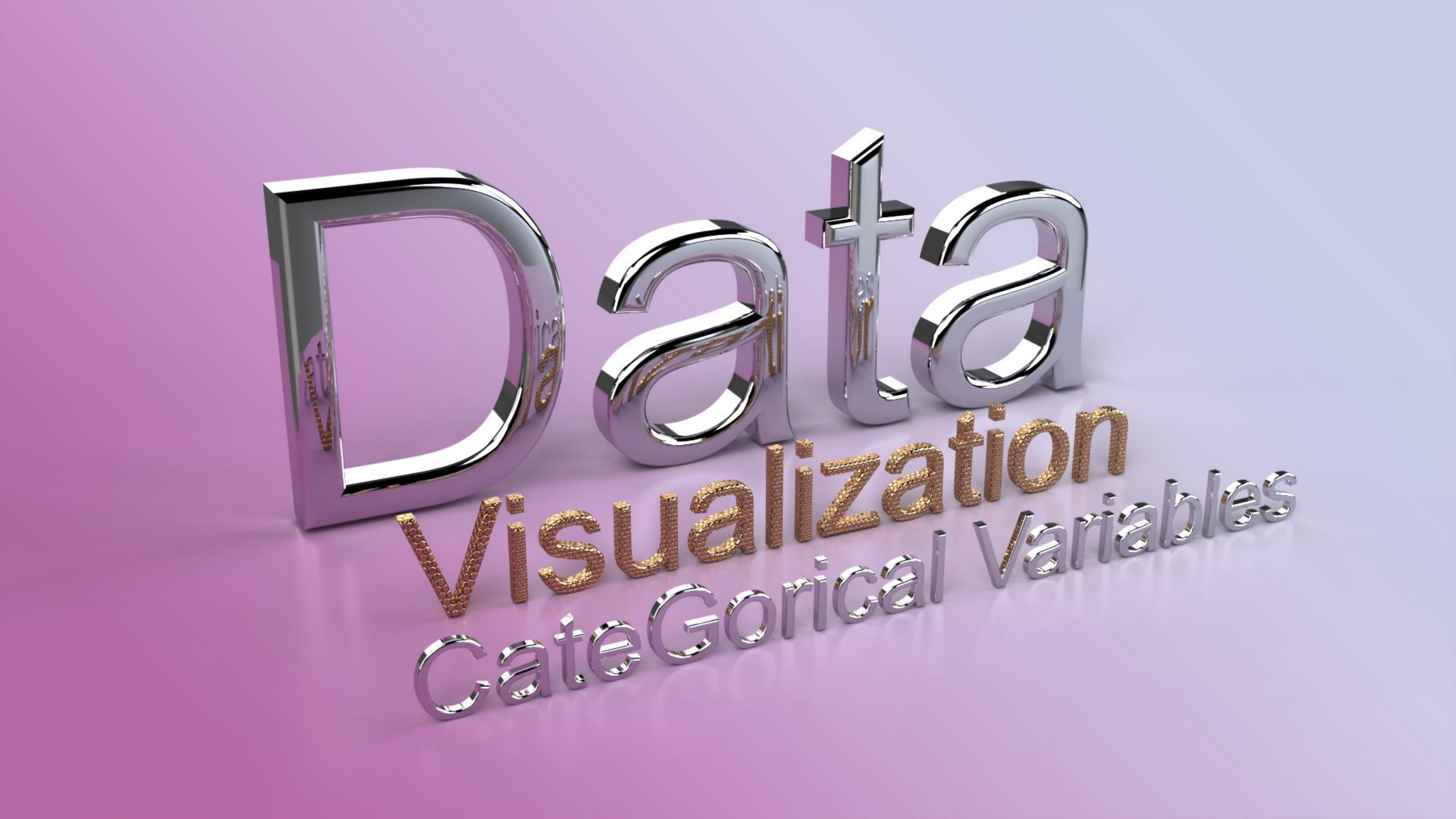

Converted Vs. Page Views Per Visit

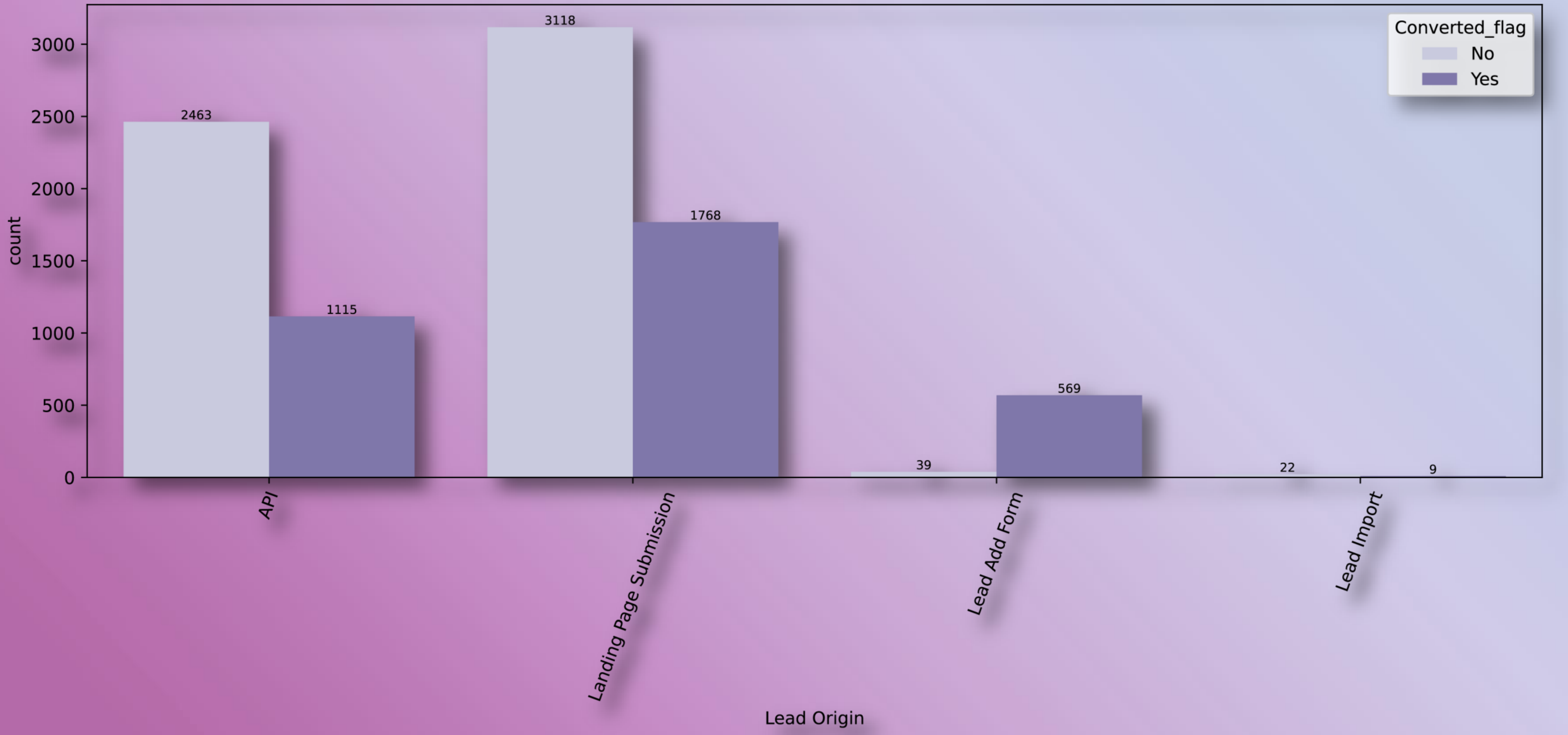

Converted Vs. TotalVisits

Median are same in both the cases.

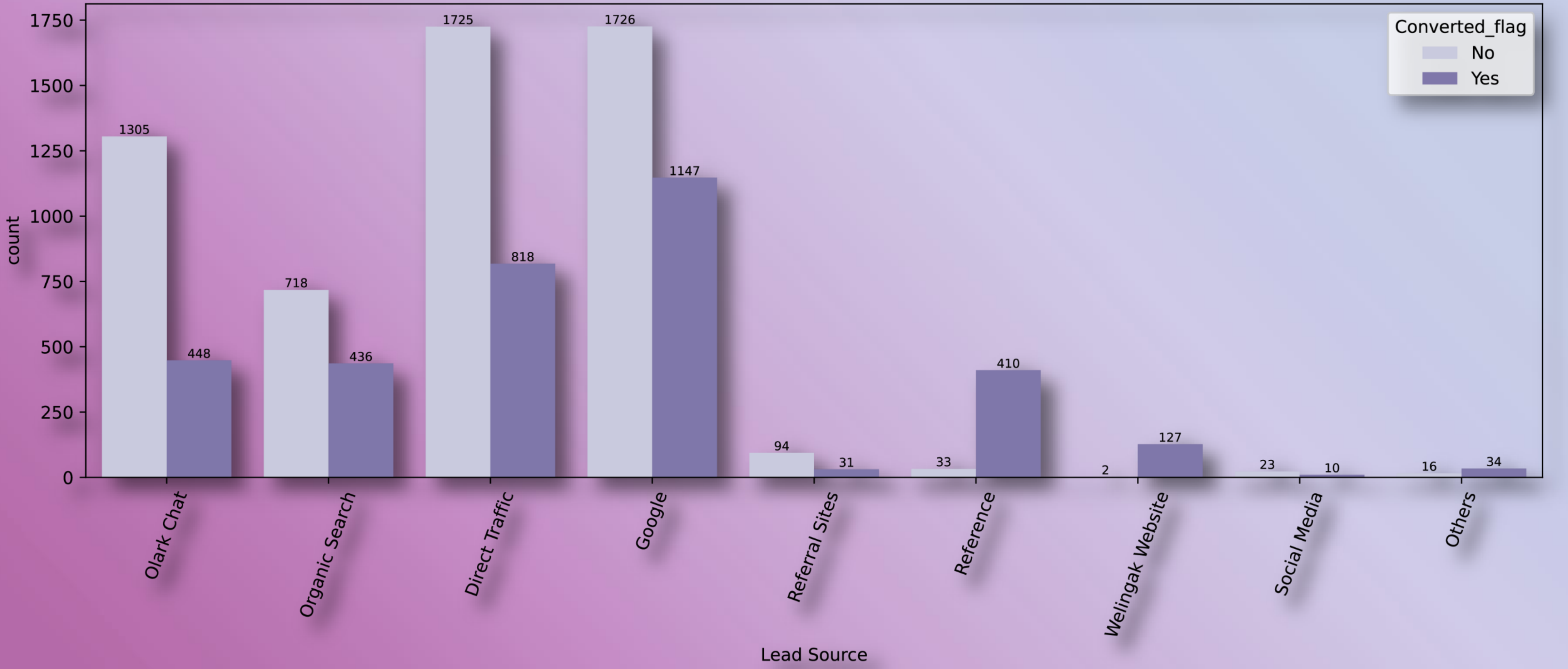

Converted Vs. Total Time Spent on Website

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

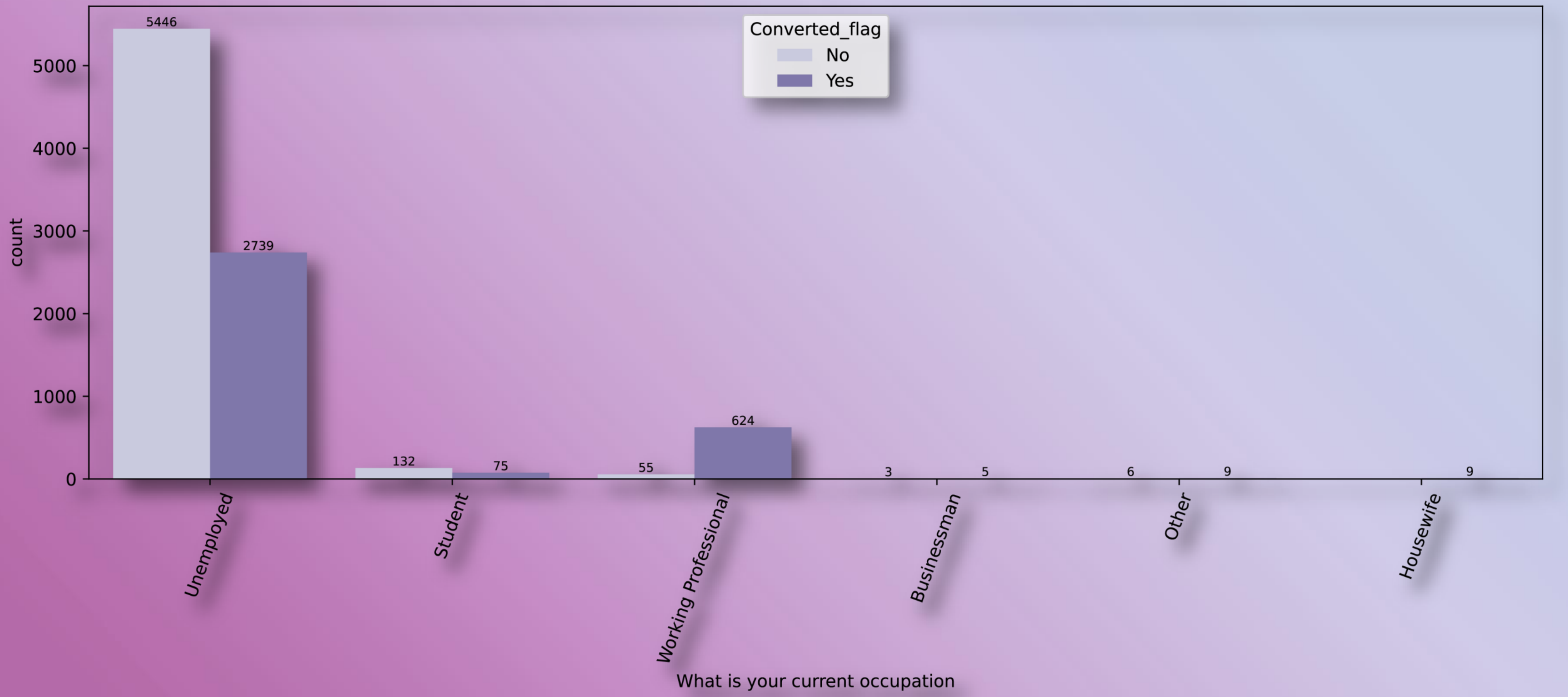

Data
Visualization
CateGorical Variables



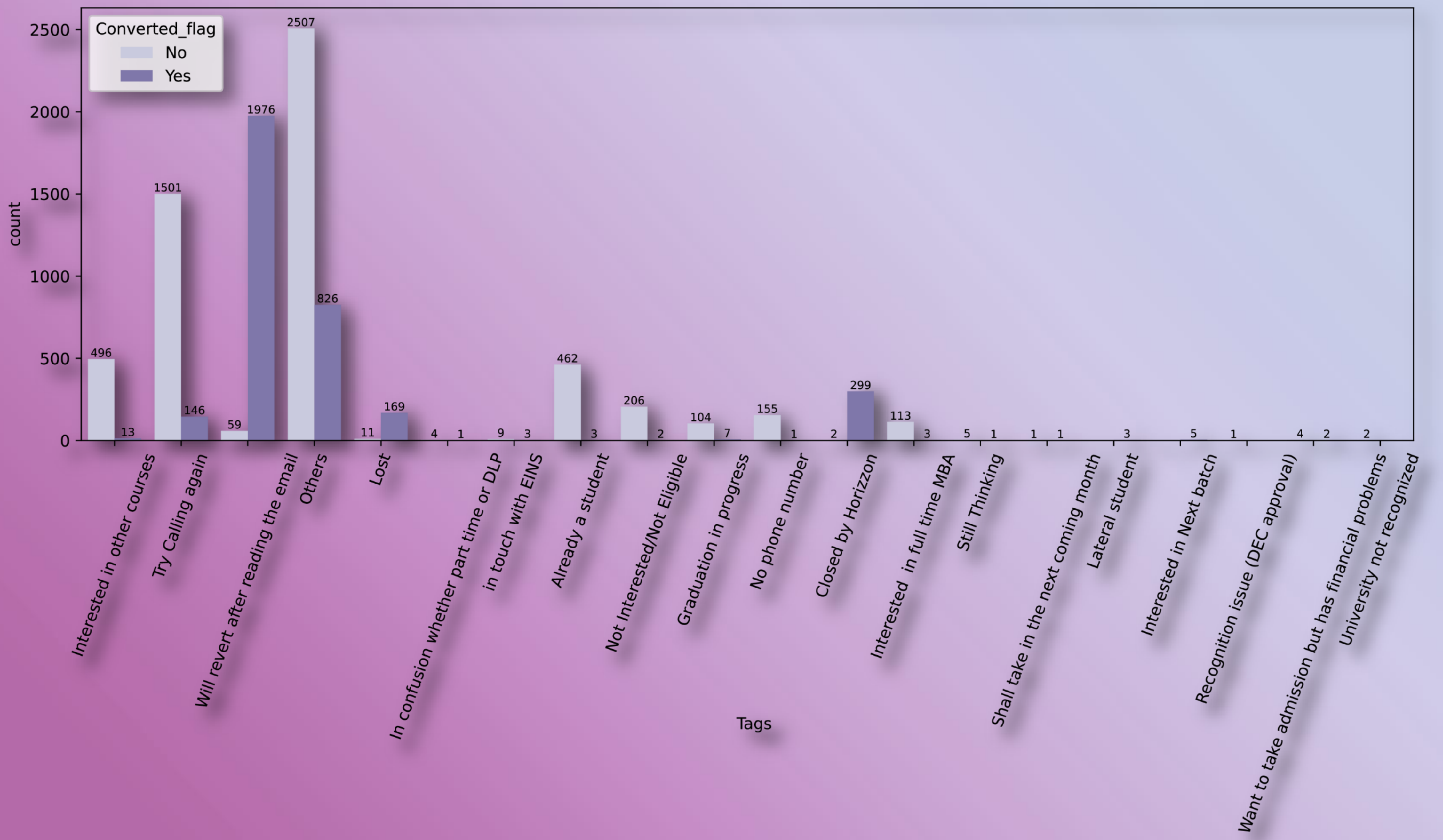
- ‘*Landing Page Submission*’ and ‘API’ generate the most leads but have less conversion rates, whereas ‘*Lead Add Form*’ generates less leads but conversion rate is great.
- Should increase conversion rate for ‘API’ and ‘Landing Page Submission’, and increase leads generation using ‘Lead Add Form’.



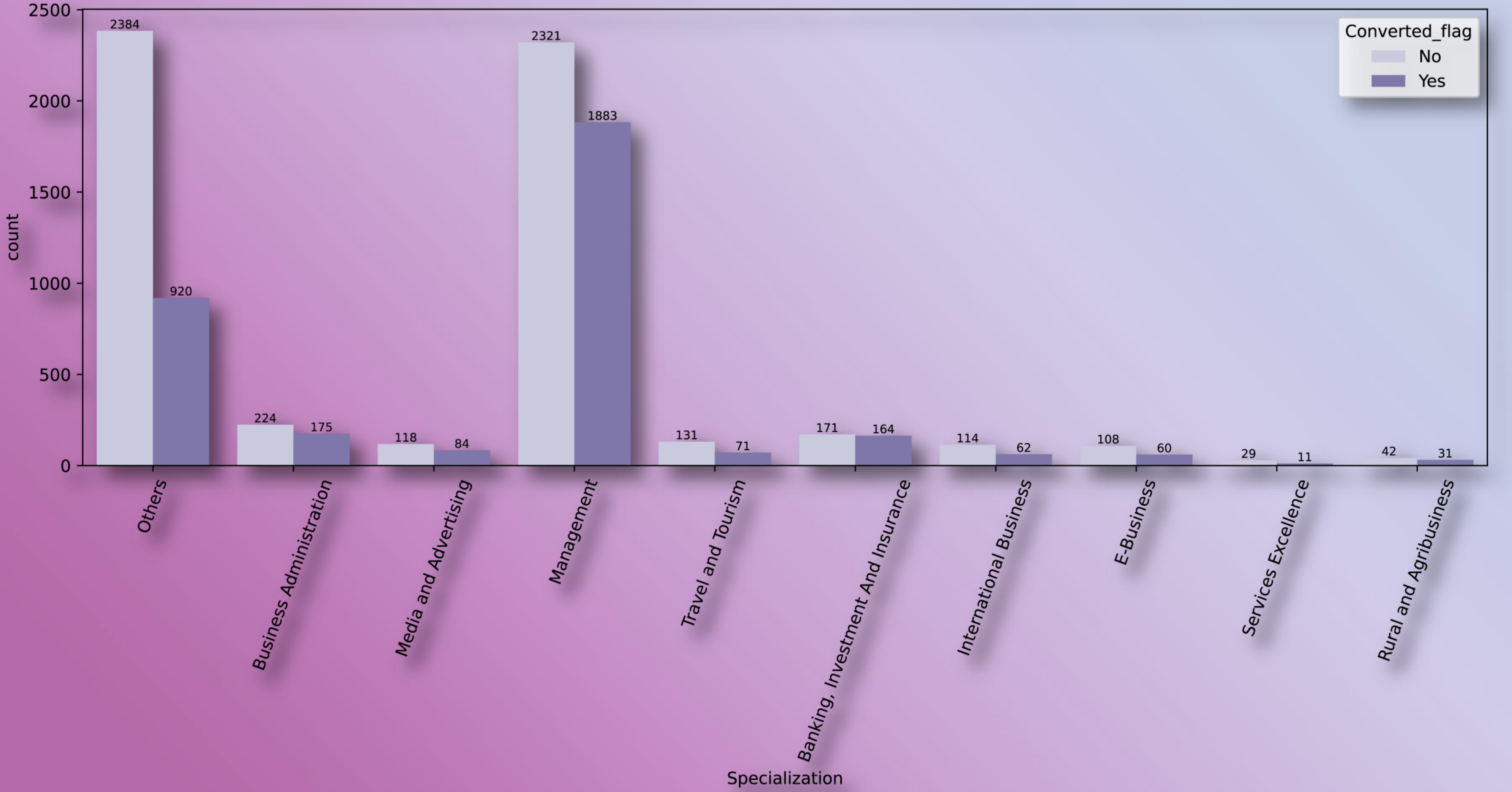
- ‘Reference’ and ‘Welingak Website’ in Lead Source has very high conversion rate.
- Most leads are generated through ‘Google’ and ‘Direct Traffic’.



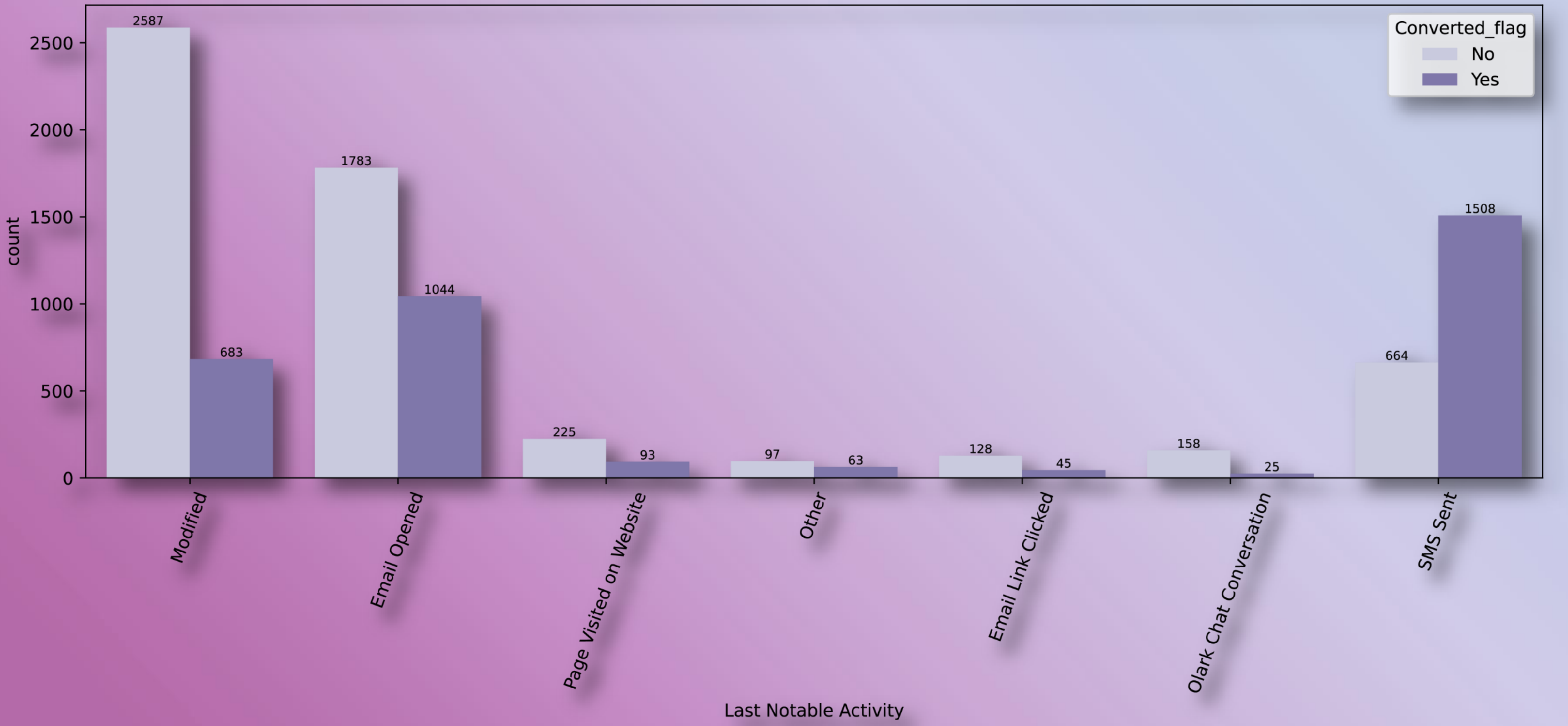
- Working Professionals are most likely to get converted.



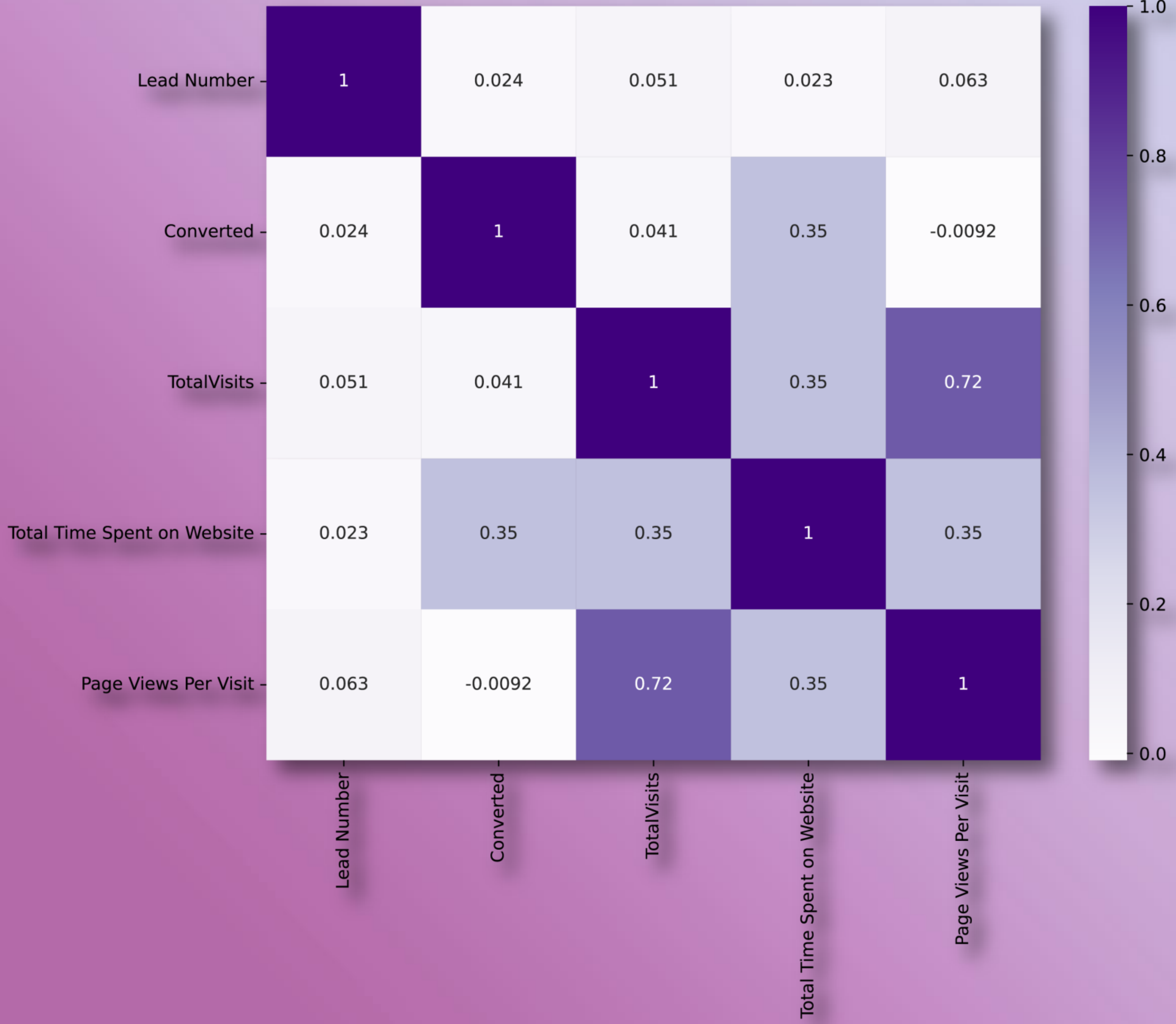
- High conversion rates for tags ‘Will revert after reading the email’, ‘Closed by Horizon’, ‘Lost to EINS’, and ‘Busy’.



- ‘Management’ in Specialization has very high conversion rate.



- Highest conversion rate is for the last notable activity ‘SMS Sent’.

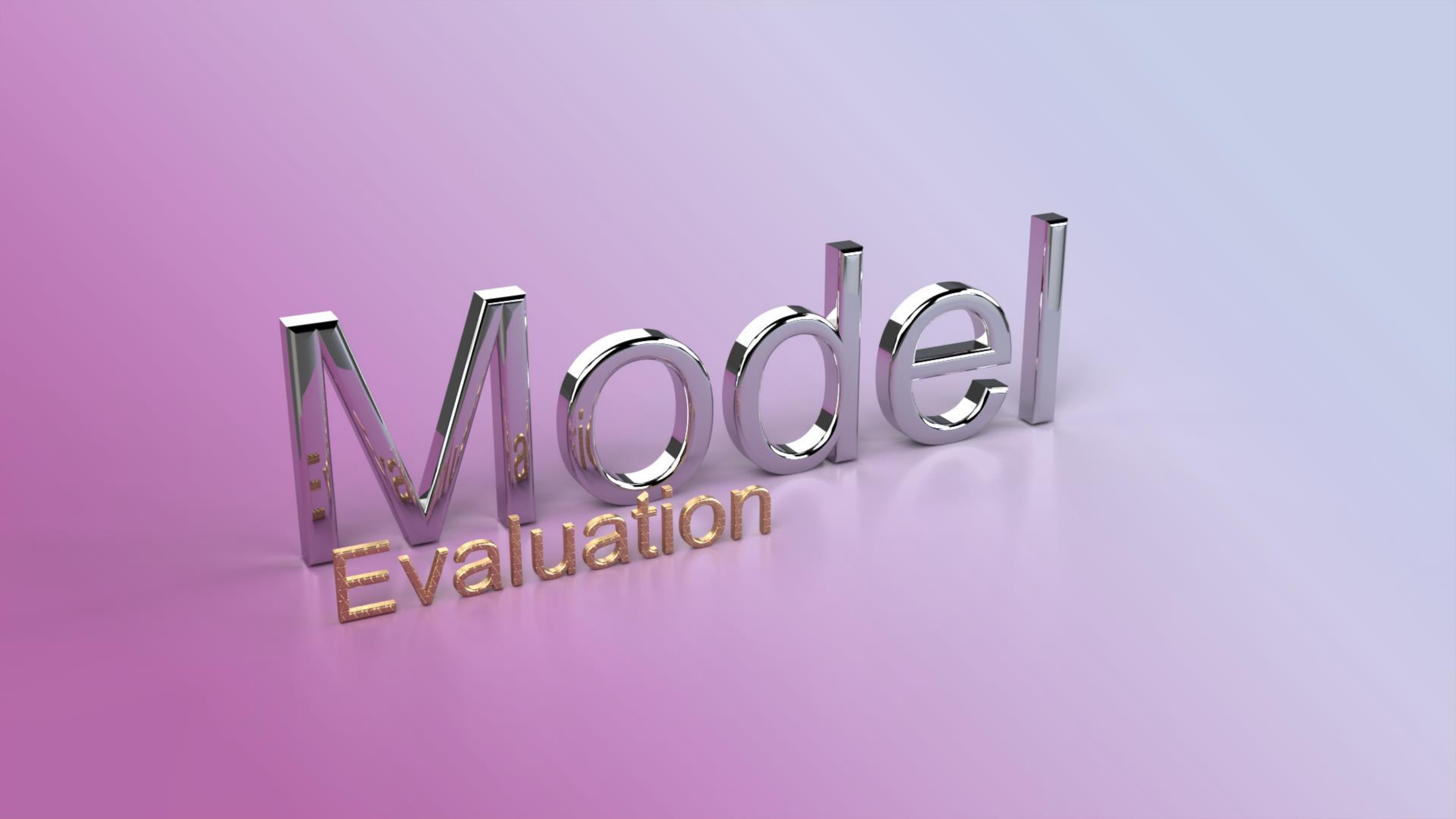


	Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit
Lead Number	1	0.024	0.051	0.023	0.063
Converted	0.024	1	0.041	0.35	-0.0092
TotalVisits	0.051	0.041	1	0.35	0.72
Total Time Spent on Website	0.023	0.35	0.35	1	0.35
Page Views Per Visit	0.063	-0.0092	0.72	0.35	1

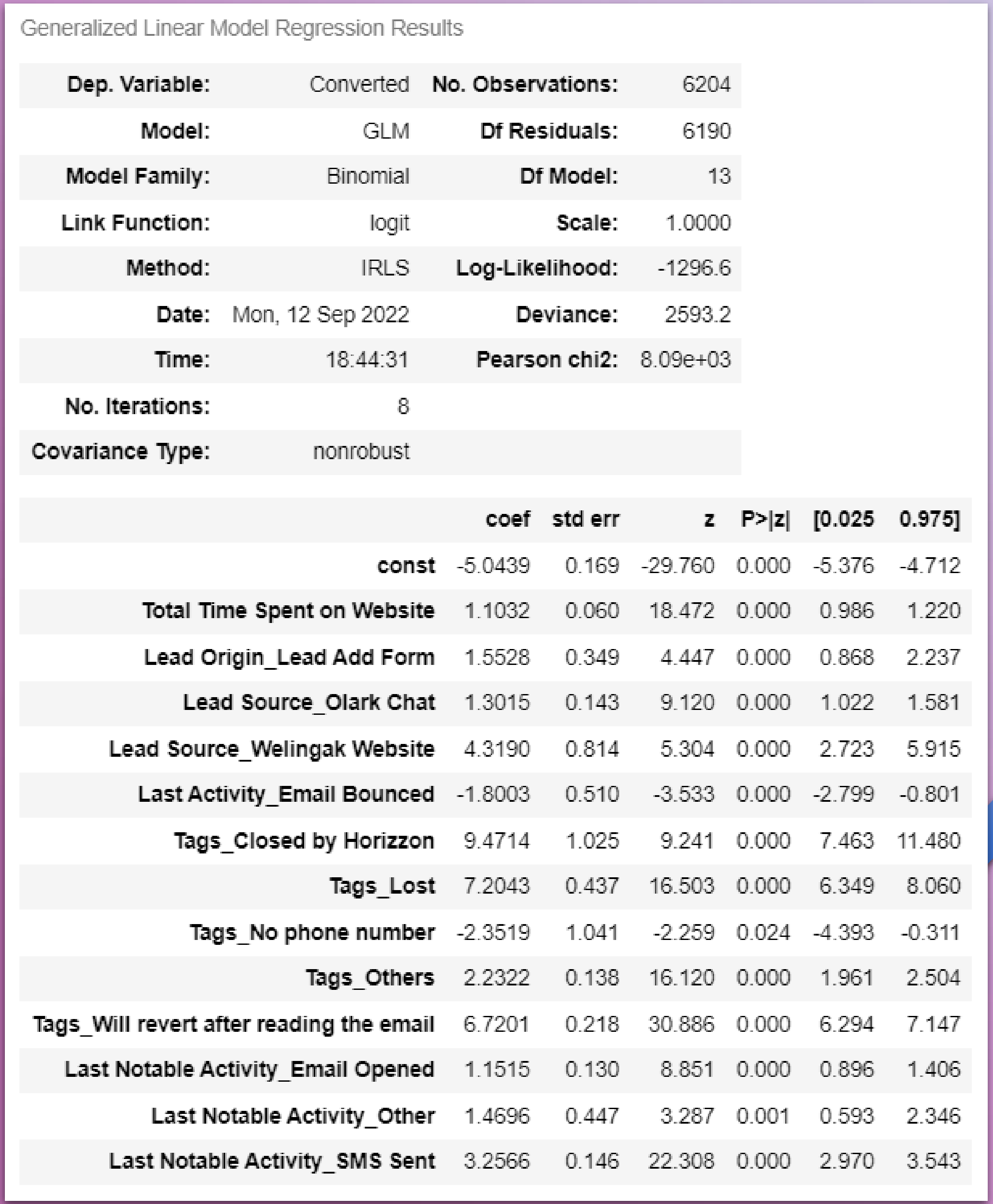
Correlation

We can observe that the variables are not highly correlated with each other.

But still there is multicollinearity among some columns.


Model
Evaluation

Final Model Summary

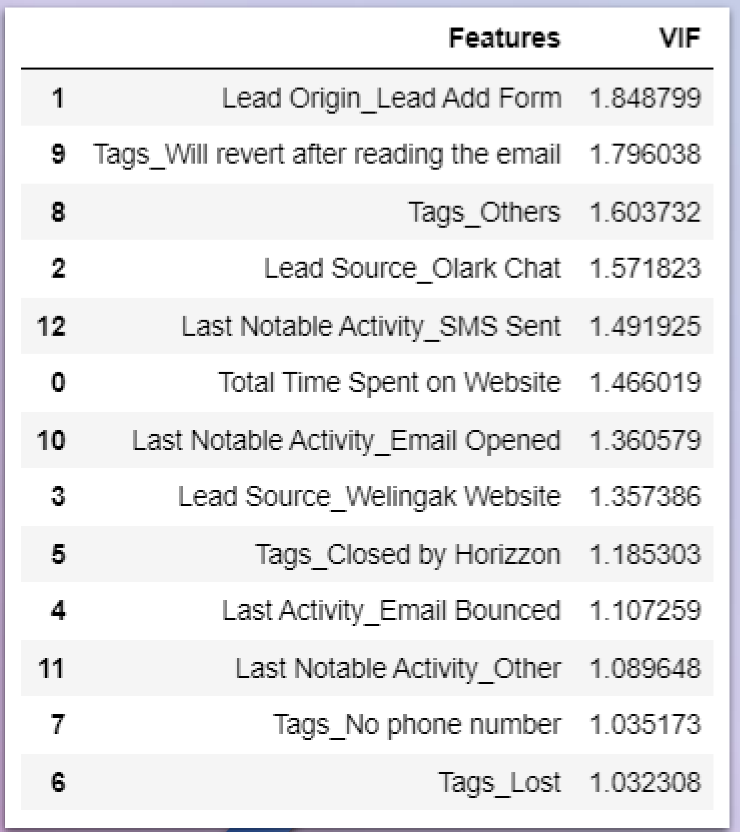


Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6204
Model:	GLM	**Df Residuals:**	6190
Model Family:	Binomial	**Df Model:**	13
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1296.6
Date:	Mon, 12 Sep 2022	**Deviance:**	2593.2
Time:	18:44:31	**Pearson chi2:**	8.09e+03
No. Iterations:	8		
Covariance Type:	nonrobust		

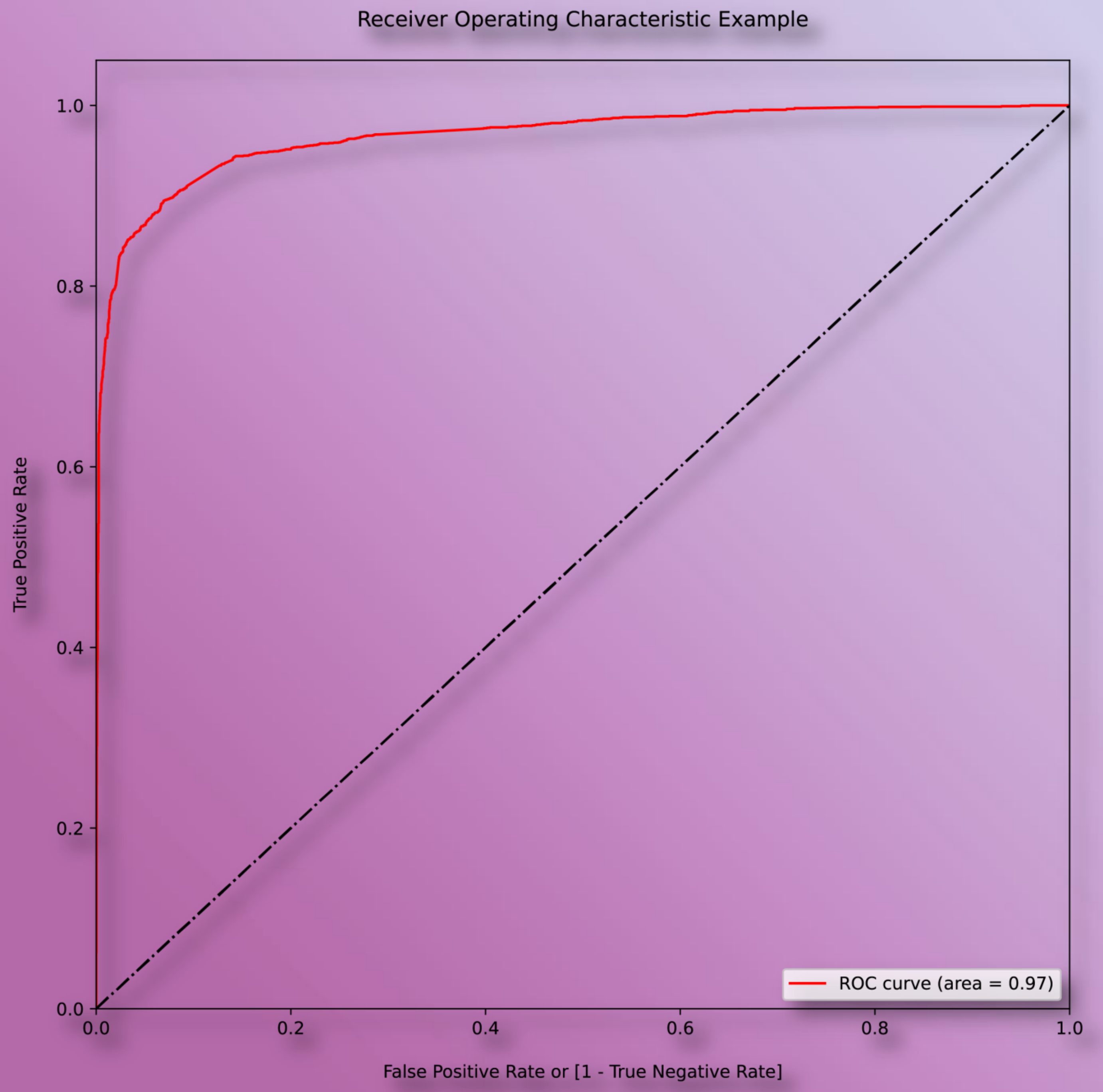
	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	-5.0439	0.169	-29.760	0.000	-5.376	-4.712
Total Time Spent on Website	1.1032	0.060	18.472	0.000	0.986	1.220
Lead Origin_Lead Add Form	1.5528	0.349	4.447	0.000	0.868	2.237
Lead Source_Olark Chat	1.3015	0.143	9.120	0.000	1.022	1.581
Lead Source_Welingak Website	4.3190	0.814	5.304	0.000	2.723	5.915
Last Activity_Email Bounced	-1.8003	0.510	-3.533	0.000	-2.799	-0.801
Tags_Closed by Horizzon	9.4714	1.025	9.241	0.000	7.463	11.480
Tags_Lost	7.2043	0.437	16.503	0.000	6.349	8.060
Tags_No phone number	-2.3519	1.041	-2.259	0.024	-4.393	-0.311
Tags_Others	2.2322	0.138	16.120	0.000	1.961	2.504
Tags_Will revert after reading the email	6.7201	0.218	30.886	0.000	6.294	7.147
Last Notable Activity_Email Opened	1.1515	0.130	8.851	0.000	0.896	1.406
Last Notable Activity_Other	1.4696	0.447	3.287	0.001	0.593	2.346
Last Notable Activity_SMS Sent	3.2566	0.146	22.308	0.000	2.970	3.543

All P-Values are less than (<0.05)



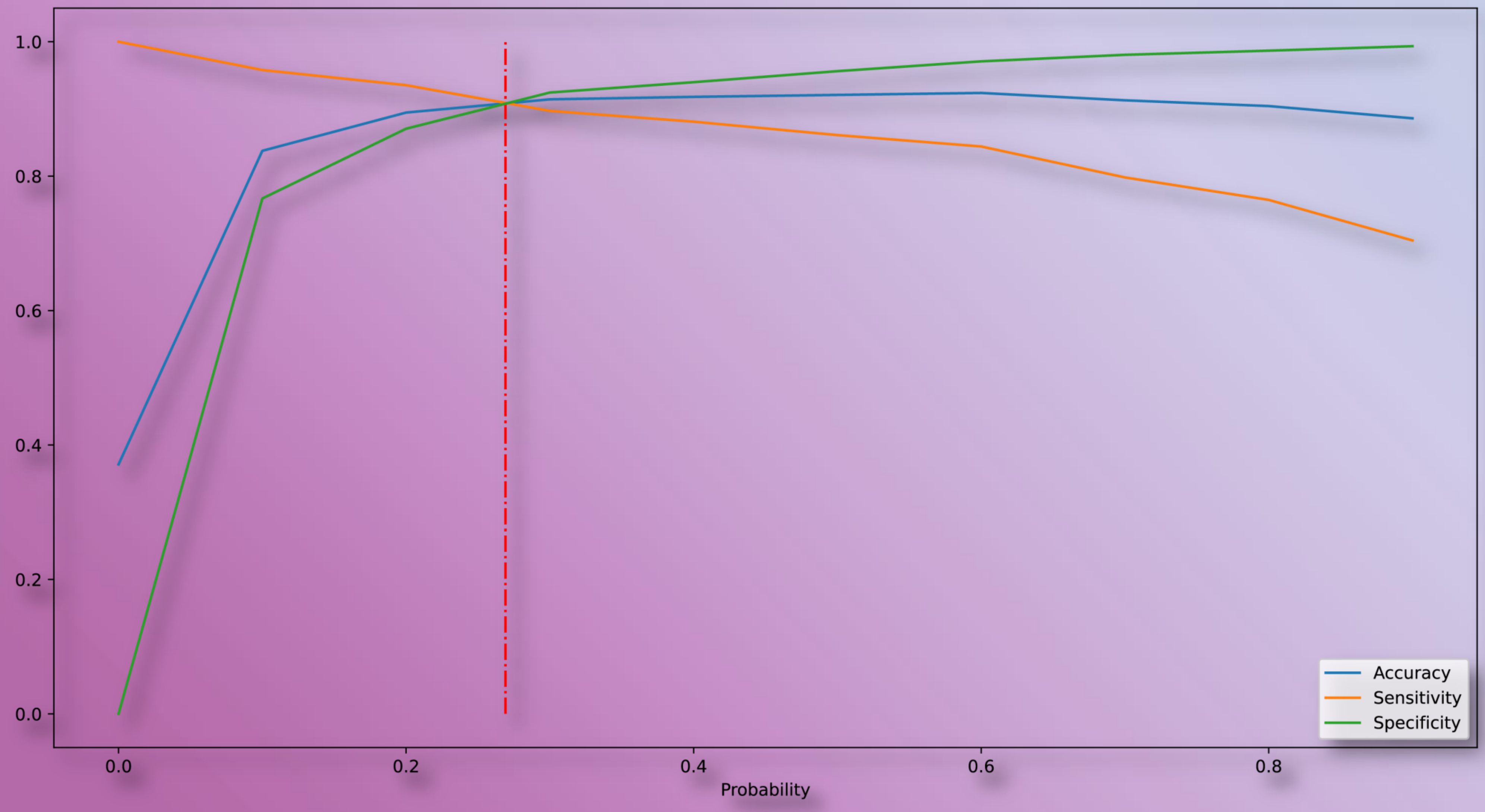
	Features	**VIF**
1	Lead Origin_Lead Add Form	1.848799
9	Tags_Will revert after reading the email	1.796038
8	Tags_Others	1.603732
2	Lead Source_Olark Chat	1.571823
12	Last Notable Activity_SMS Sent	1.491925
0	Total Time Spent on Website	1.466019
10	Last Notable Activity_Email Opened	1.360579
3	Lead Source_Welingak Website	1.357386
5	Tags_Closed by Horizzon	1.185303
4	Last Activity_Email Bounced	1.107259
11	Last Notable Activity_Other	1.089648
7	Tags_No phone number	1.035173
6	Tags_Lost	1.032308

All VIF values are less than 5



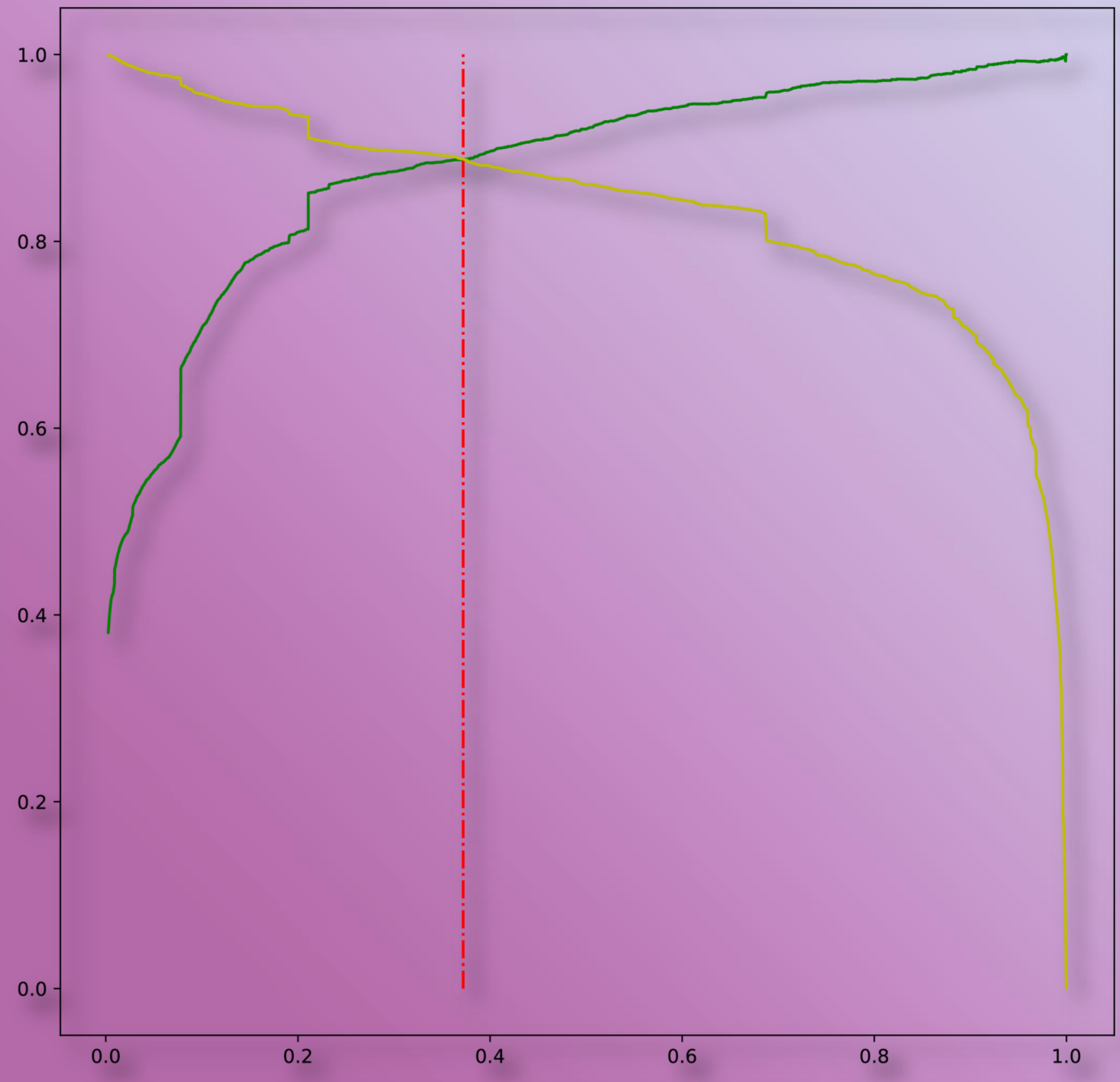
Area Under ROC Curve : 0.97

Optimal Threshold



Optimal Cut-off = 0.27

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



Precision or Recall Tradeoff

Recall Tradeoff Intersection Point : 0.372

Model Matrix

Train Data

- Confusion Matrix :

	Non-Converted Leads	Converted Leads
Non-Converted Leads	3590	311
Converted Leads	231	2072

- Accuracy : 91%
- Sensitivity : 90%
- Specificity : 92%
- Precision : 87%
- Recall : 90%
- F1-Score : 88%

Model Matrix

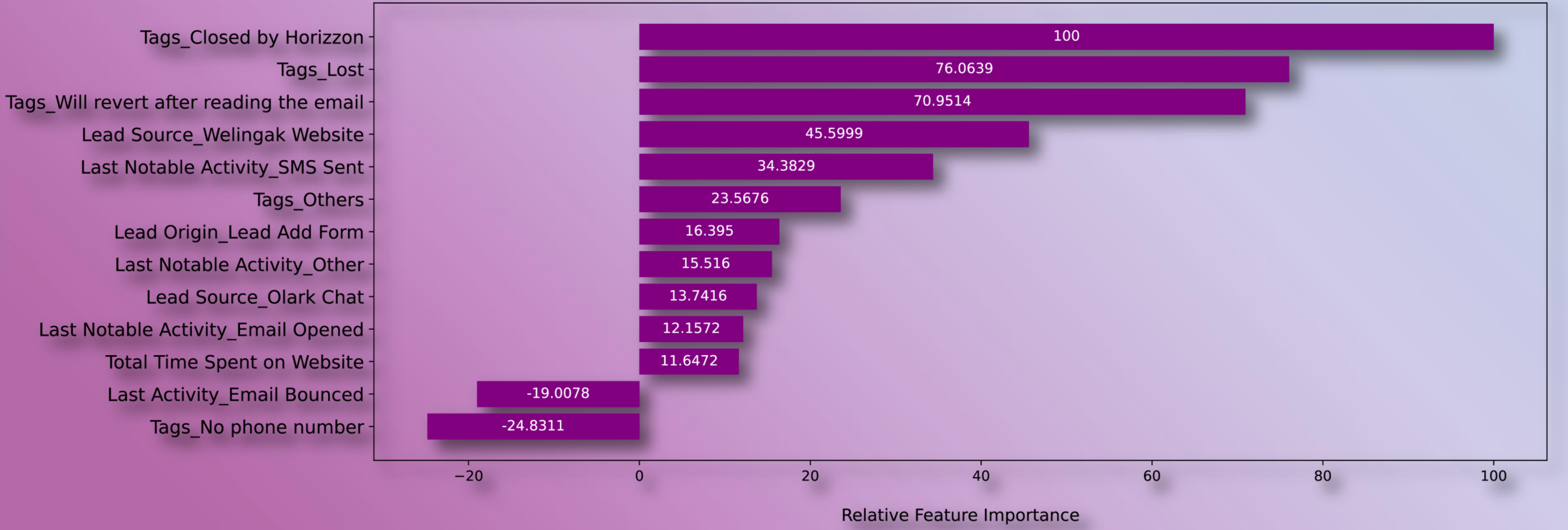
Test Data

- Confusion Matrix :

	Non-Converted Leads	Converted Leads
Non-Converted Leads	1533	94
Converted Leads	112	920

- Accuracy : 92%
- Sensitivity : 89.147%
- Specificity : 94.222%
- Precision : 91%
- Recall : 89%
- F1-Score : 90%

Importance Of Features



Conclusion

Focus :

Company should focus on following features to increase the leads

- **Tags_Lost :** Leads that have been tagged as ‘Lost ‘also contribute to the conversion to a considerable extent.
- **Tags_Closed by Horizzon :** Leads that have been assigned Tags as ‘closed by horizon’ have the highest probability of conversion.
- **Tags_Will revert after reading the email :** Leads that have been tagged as ‘will revert after reading the mail’ also have significant correlation with the conversion.

Expansion:
Company should also focus on Lead Score (which are the probabilities obtained via algorithm) which are greater than 80% to expedite the conversion rate.