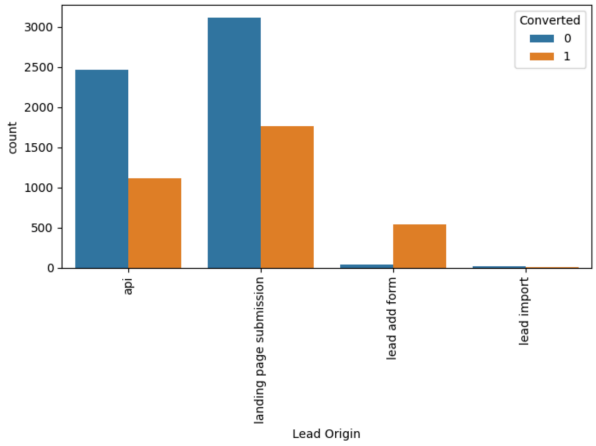
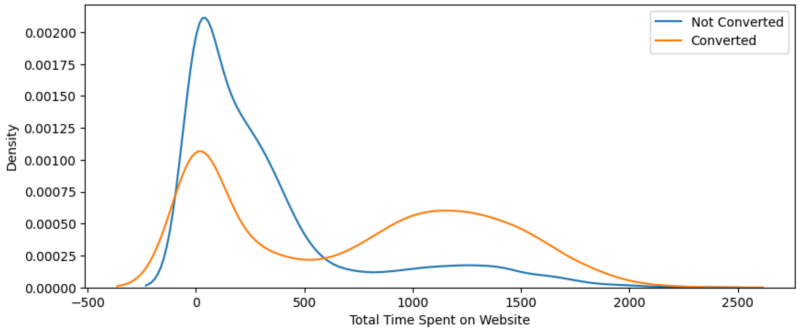
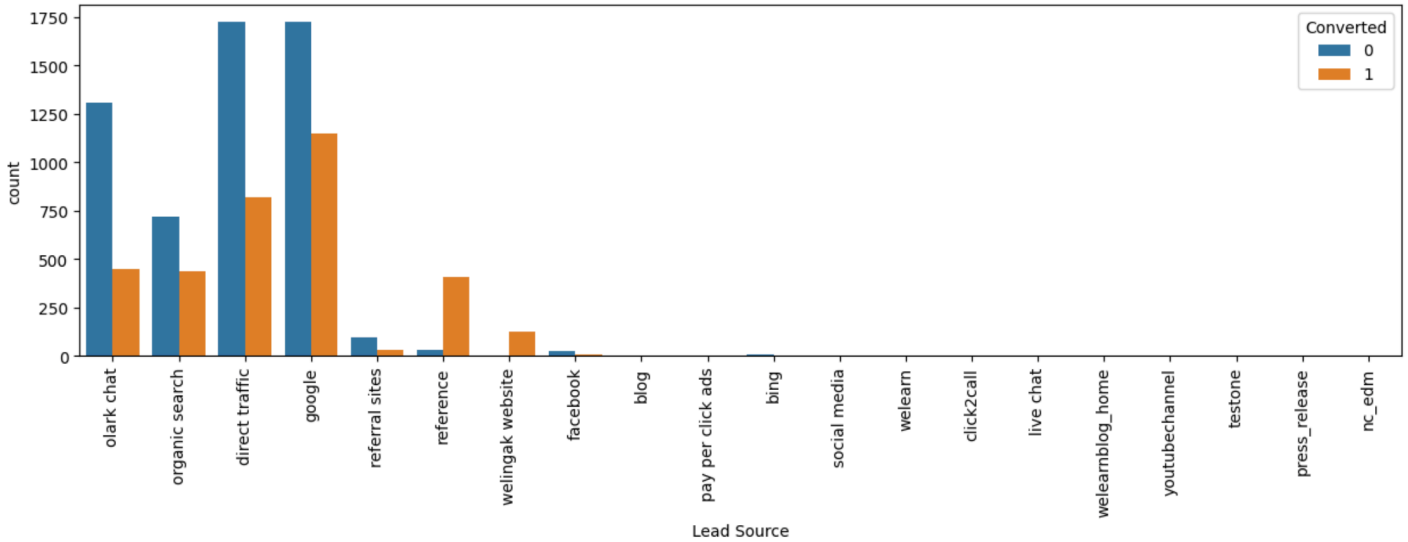
Summary Report

Our goal was to understand the problem of the company, visualise their data to get insights in order to make recommendations and build a Logistic Regression model that would be able to give a lead score between 0 to 100 depending on the chances of that lead converting successfully.
We started with a dataset of around 9000 data points each with 36 features ranging from where the lead originated from, what was the source, how much time the person spent on the website, how many pages he/she visited, their country and city and their asymmetric activity index etc. We were given a data dictionary explaining what those features meant.
We started with cleaning the data so that it is in the right form before we start making visualisations and build our model. There were quite a lot of features that had to be removed because of missing values.
One example of that was the 'How did you hear about X Education' column which had 78% values missing. We had to drop it.
In some cases, we had only few missing values such as in 'Lead Source', only 0.38% of the values were missing. We just dropped those rows and kept this column.
While in other cases, we actually imputed some value in the place of missing values. One such column was 'Specialization'. 36% of values were missing. But they were missing because the students didn't fill those. There can be multiple reasons for that. Maybe they didn't find what they were looking for. Maybe they were just not sure at that moment to make a decision. So we filled 'unstated' in those missing values.
After tackling all the missing values in our dataset. We continued with cleaning our dataset. Such as we had to drop a column if it was highly imbalanced or in other words, it didn't have the necessary amount of variance to be conducive to our analysis or making predictions. For example: 'Country'.
After that, we started making visualisations via univariate and bivariate analysis. Here are some examples of that:







Here are some conclusions we made from the visualisations shown here:

Lead Origin
Try improving the conversion ratio in the 'api' and 'landing page submission' categories
Take advantage of the leads coming from 'lead add form' and try to increase the numbers

Total Time Spent on Website
For values above 500 (approximately), we can see that leads do convert
It makes sense since people who spent more time on the website were probably serious about the product

Lead Source
Try to improve the conversion rate if possible in the 'olark chat', 'organic search', 'direct traffic', 'google' categories
Take advantage (i.e. get more number of leads) of the 'reference' and 'welingak website' categories, they perform remarkably well when it comes to conversion

After creating dummy variables out of categorical columns, we divided the dataset into Train (70%) and Test (30%) set and scaled the numerical features. We used Recursive Feature Elimination to get the best features to be used for model building. Then we build our Logistic Regression model using statsmodels library. We used the p-value and variance inflation factor to remove the features which were not needed the best. And after that, we used various accuracy metrics like accuracy, confusion matrix, specificity, sensitivity, precision and recall on both Train and Test sets with the thresholds giving us the best values. Here is a summary of the final model we built:-

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6246
Model:	GLM	Df Residuals:	6233
Model Family:	Binomial	Df Model:	12
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2626.9
Date:	Mon, 22 May 2023	Deviance:	5253.8
Time:	16:43:53	Pearson chi2:	6.30e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.3882
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.4387	0.124	-3.538	0.000	-0.682	-0.196
Total Time Spent on Website	1.1131	0.040	27.692	0.000	1.034	1.192
Lead_Origin__landing page submission	-1.0646	0.129	-8.282	0.000	-1.317	-0.813
Lead_Origin__lead import	1.2644	0.482	2.622	0.009	0.319	2.209
Lead_Source__olark chat	1.1864	0.124	9.604	0.000	0.944	1.429
Lead_Source__reference	3.3094	0.239	13.874	0.000	2.842	3.777
Lead_Source__welingak website	5.8156	0.730	7.968	0.000	4.385	7.246
Last_Activity__email bounced	-2.4092	0.377	-6.387	0.000	-3.148	-1.670
Last_Activity__olark chat conversation	-1.4864	0.169	-8.790	0.000	-1.818	-1.155
Specialization__unstated	-1.1158	0.125	-8.954	0.000	-1.360	-0.872
What_occupation__working professional	2.6397	0.194	13.622	0.000	2.260	3.019
Last_Notable_Activity__others	1.1011	0.272	4.045	0.000	0.568	1.635
Last_Notable_Activity__sms sent	1.5909	0.080	19.994	0.000	1.435	1.747