Lead Scoring Case Study

Conducted by: Khushboo Mudgal
Atimukta Ghosh
Ramya Vedantam

Problem Statement

The education company, X Education, currently has a lead conversion rate of 30%. It aims to increase the rate to 80%, by identifying the hot leads using a data science approach, and especially targeting those leads with the help of the sales team.

Note 1: A lead is defined as a person who had landed on the company's website, and filled up a form by providing the email ID or phone number.

Note 2: A hot lead is a lead who has a greater potential of conversion.

Note 3: Conversion happens when a lead becomes a paying customer of the company.

Approach used

The initial cleaning of the data included handling features with high proportion of missing values and other inconsistencies, removing cases with high proportion of missing values, null values, and null-like values (such as the “Select” entry indicating no relevant option was chosen), and imputing or dropping remaining missing values.

The EDA stage included assigning dummy variables to the categorical variables, exploring univariate characteristics of the numerical variables, and exploring bivariate relationships of the target variable, Converted, with the numerical variables.

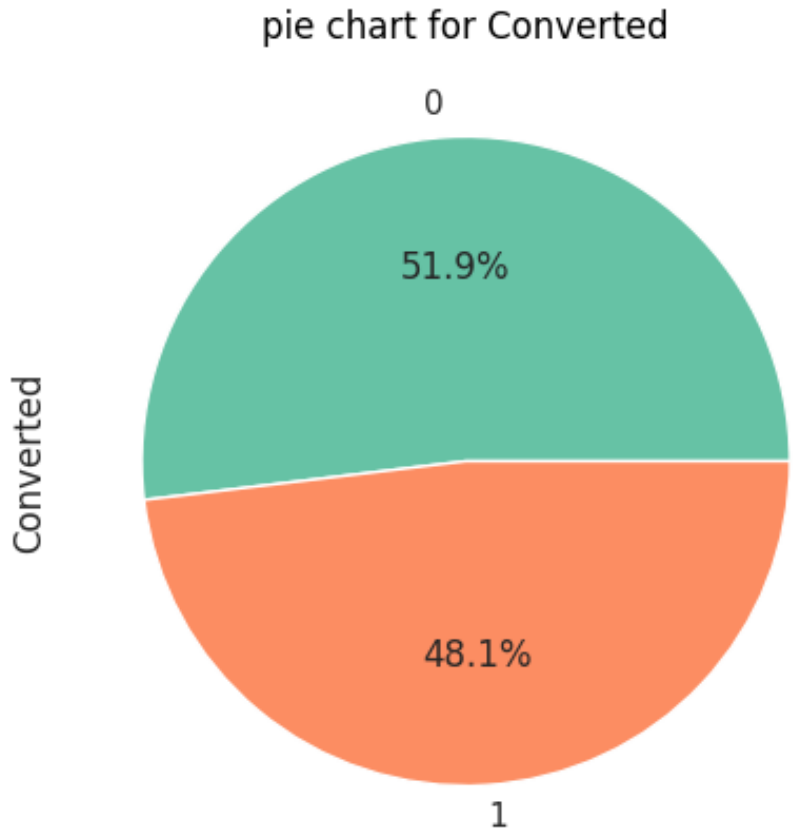
A logistic regression analysis is used, since the target is categorical in nature, and the explanatory variables are a mixture of numerical and categorical variables.

Data Cleaning: Missing Values Treatment

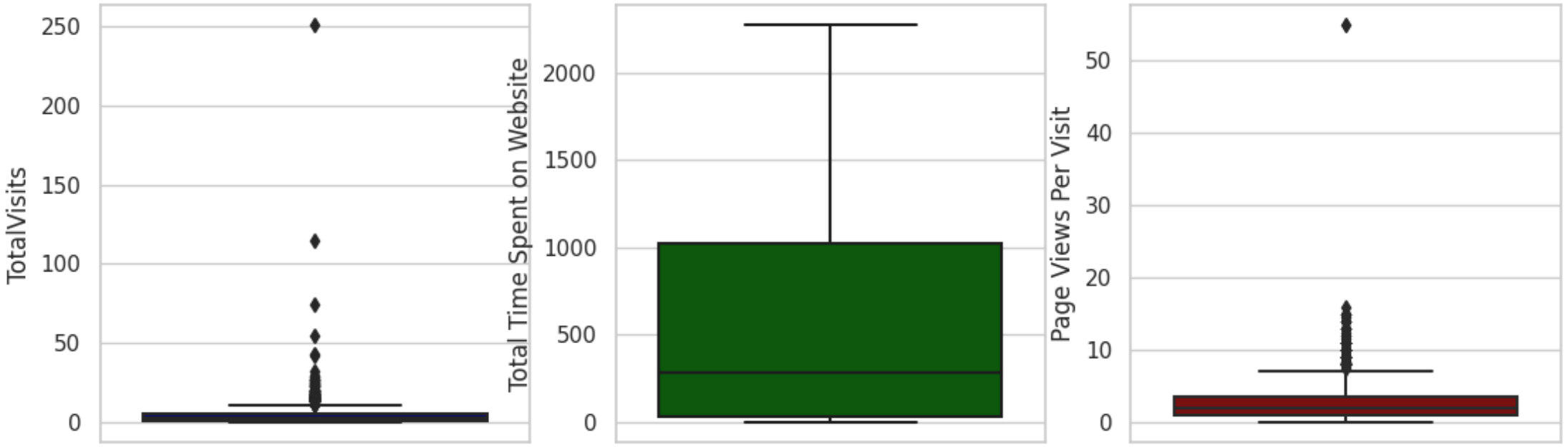
- The original dataset had 37 column, including the target variable, Converted.
- After dropping the columns with more than 35% missing values, the column of ID and lead number, we are left with 12 columns.
- Of the remaining columns, 8 are categorical, and 3 are numerical, apart from the target column.

Target Variable in Cleaned Dataset

In the cleaned dataset, 48.1% of the leads converted (Converted=1), while the remaining 51.9% of the leads did not convert (Converted=0), as exhibited by the attached pie chart.



Distribution of Numerical Features: Boxplots

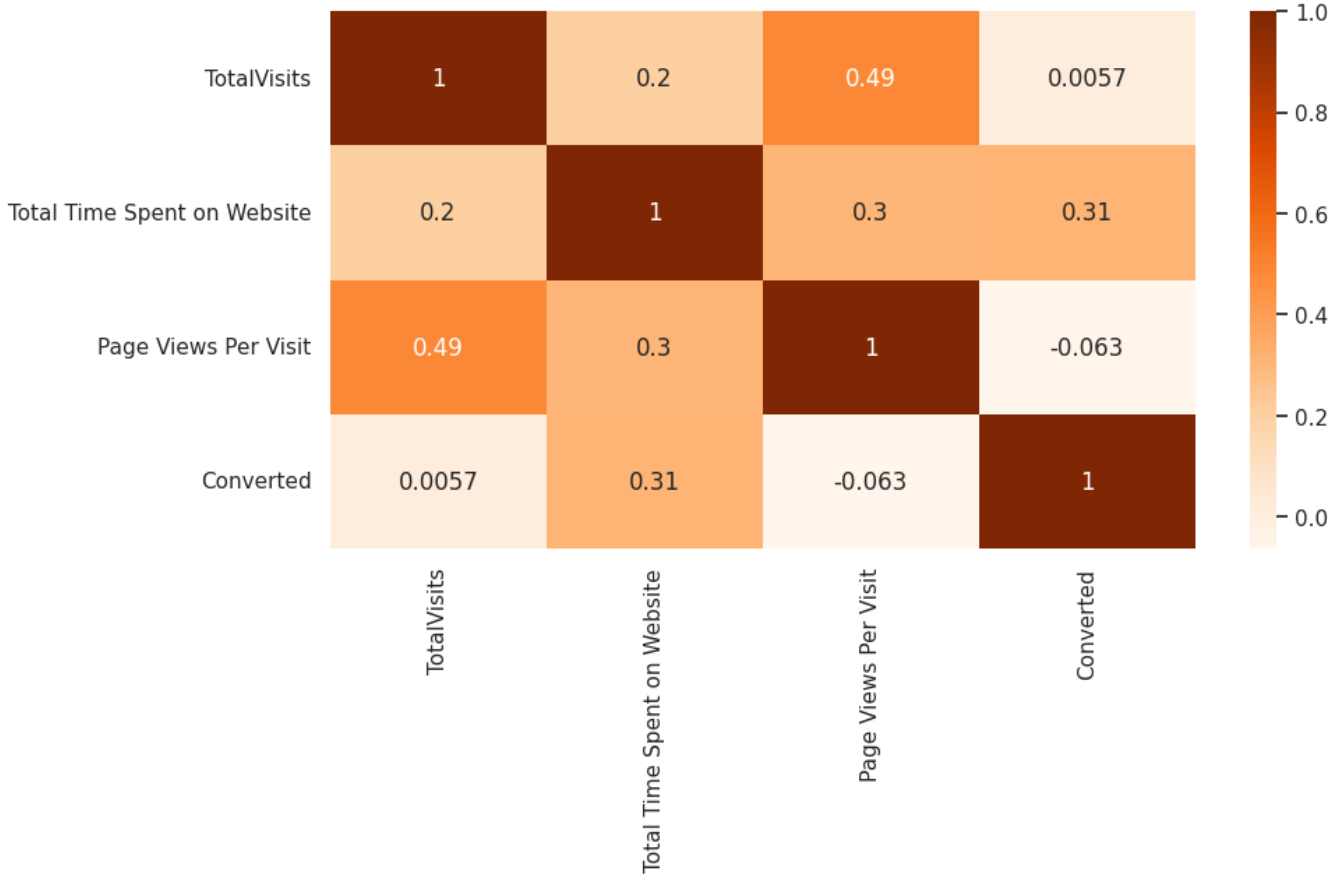


Distribution of Numerical Features

- The distribution of TotalVisits has a large number of positive outliers.
- The distribution of “Page Views Per Visit” has a large number of positive outliers.
- The distribution of “Total Time Spent on Website” does not have any outliers.
- We have proceeded with the analysis without deleting the outliers.

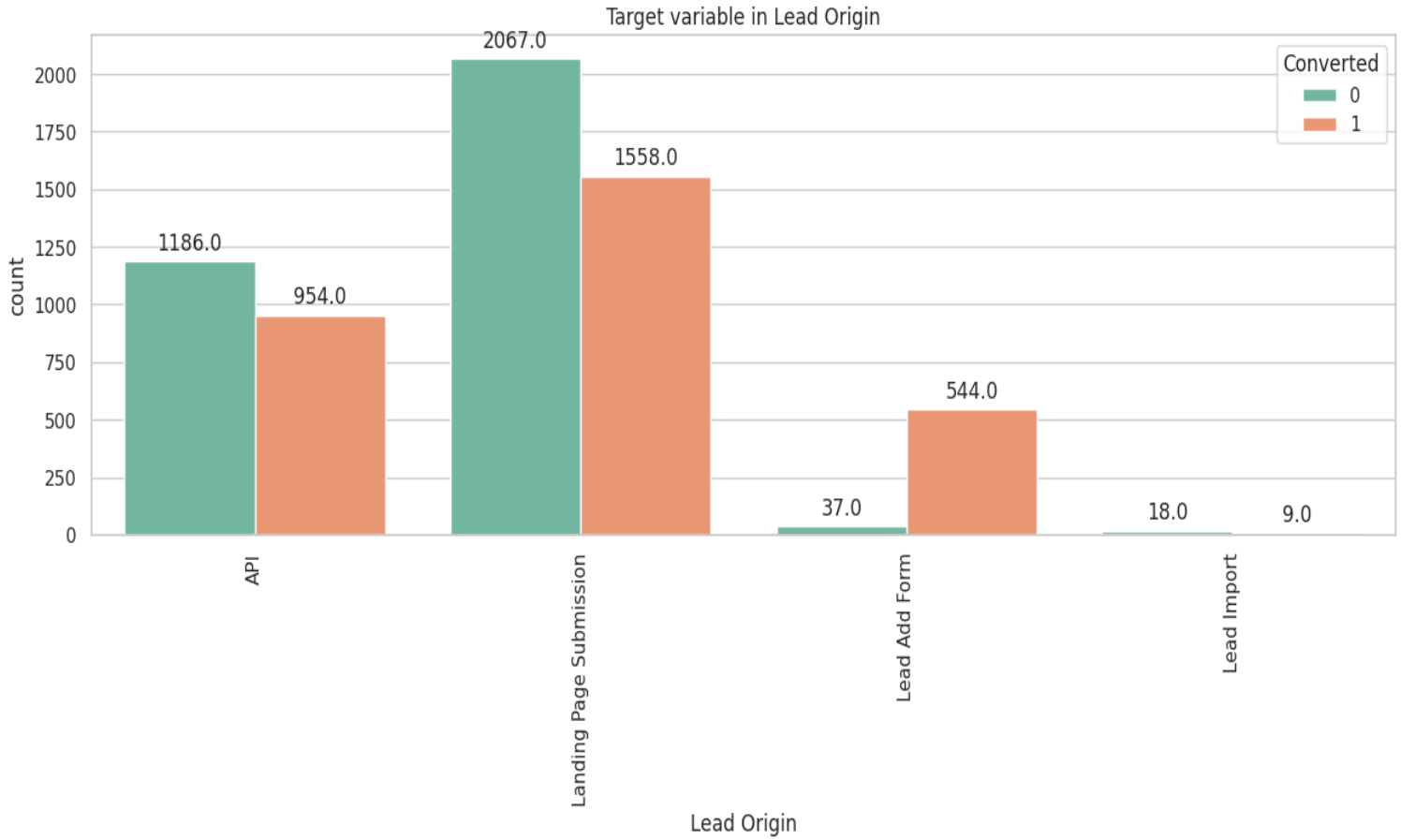
Relationship between Target and Numeric Features

The attached heatmap shows the bivariate linear correlations of the target variable with the numerical variables. It makes sense that the magnitudes of the correlations are not large, as the model being fitted is not a linear regression, but a logistic regression model.



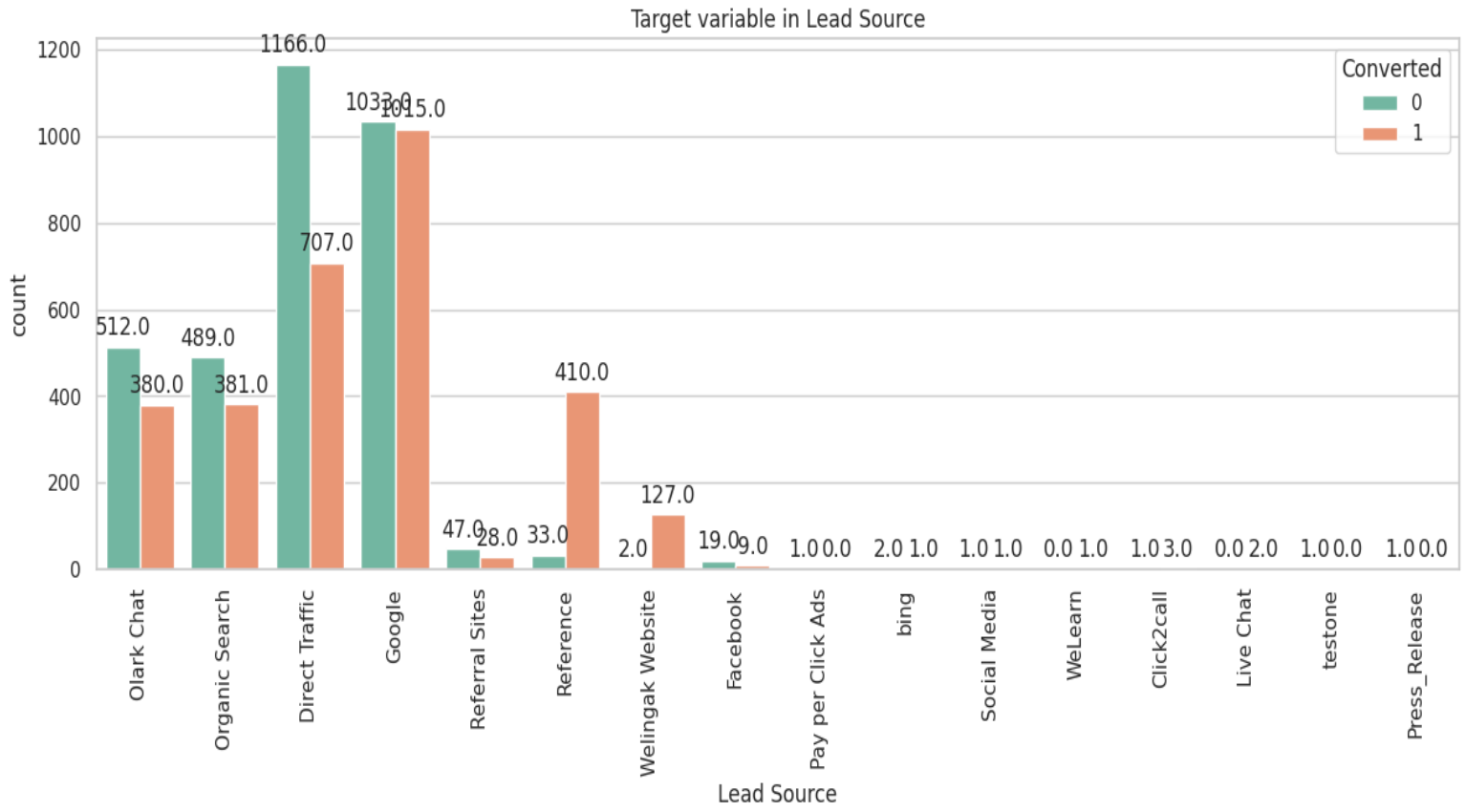
Relationship between Target and Lead Origin

Lead Origin is an important categorical feature that has helped in estimating the conversion probability. The counts of converted and not converted cases across the various lead origins are shown.



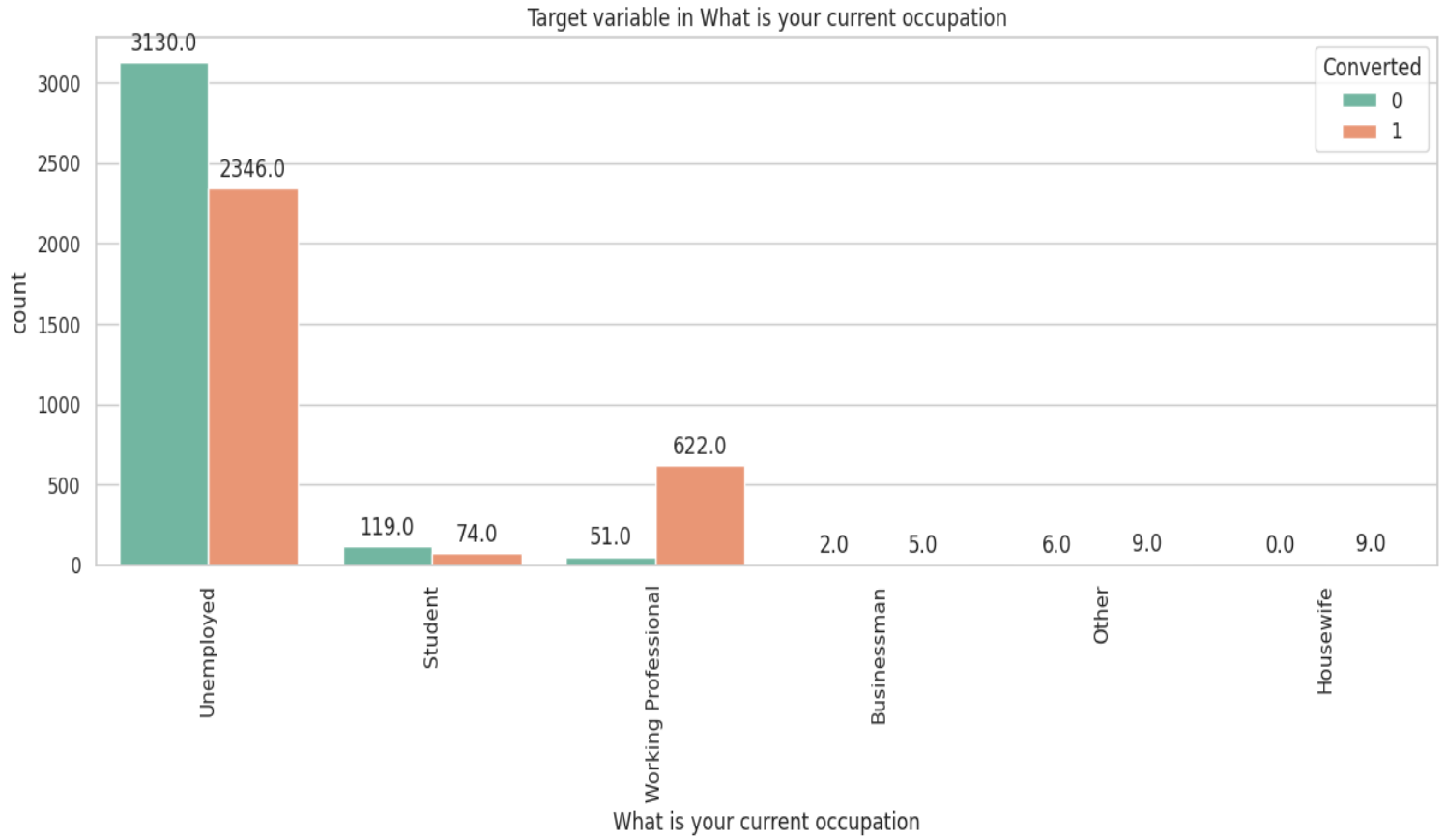
Relationship between Target and Lead Source

Lead Source is yet another important categorical feature in the estimation. The counts of converted and not converted leads across the various lead sources are shown.



Relationship between Target and Current Occupation

A lead's current occupation is one of the top three important categorical features in the model. The counts of converted and not converted leads across the various current occupations are shown.



Categorical Features: Dummy Variables

- In the cleaned dataset, dummy variables are assigned to the categorical features.
- For each categorical feature, number of dummy variables is one less than the number of categories in that feature.
- As the default number of dummy variables is equal to the number of categories of a feature, once dummy must be dropped for each categorical feature.
- The original columns of the categorical features containing the names of categories instead of the numerical values of the dummy variable are dropped to ensure no duplication takes place.

Train-Test Split

- The finally cleaned dataset is split into 70% train set, and 30% test set.
- The model must be trained using the train set, and later evaluated by applying it to the test set.
- The numerical features are scaled using MinMaxScaler(), to ensure the variations in magnitudes across the numerical feature do not distort the interpretation of the final model.

Model Building

- First, Recursive Feature Elimination (RFE) is used to extract the top 15 most important features. Note that, due to assignment of dummy variables, the number of columns in the dataset is now too large. It is illogical to include such a large number of features in a model; 15 is still a large number of features, used here as a starting point before further elimination of unnecessary features.
- The first logistic regression model is obtained using the 15 features suggested by RFE.
- The Variance Inflation Factor (VIF) is calculated for each feature as a measure of multicollinearity among features.

Refining the Model

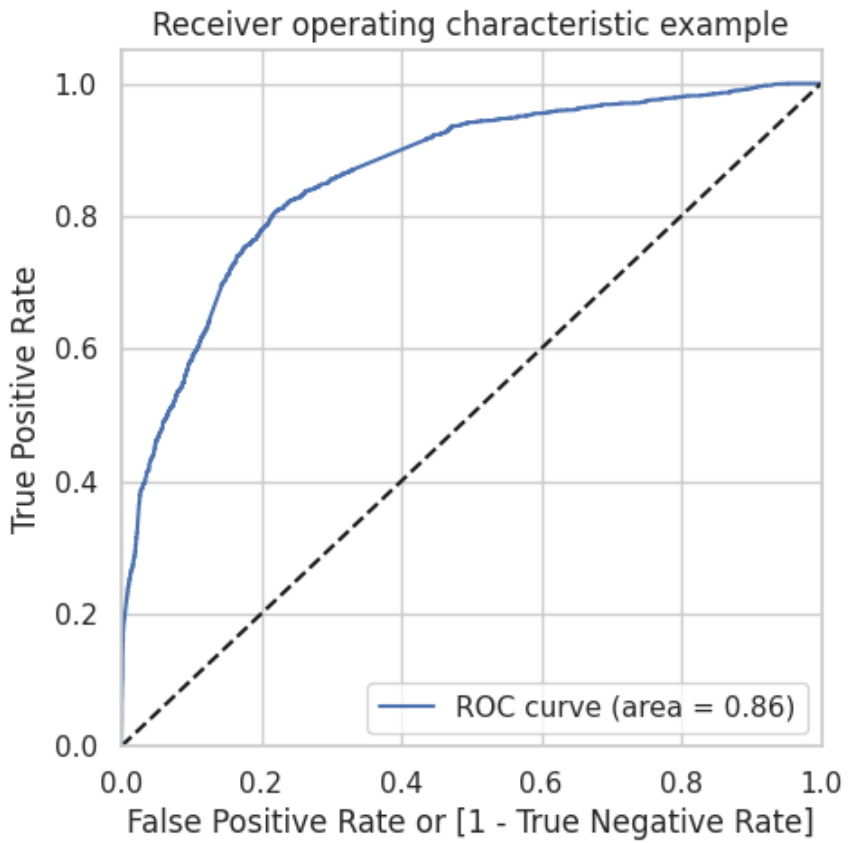
- The feature with the highest VIF value exceeding 5 is dropped. The model is fitted again with the remaining features, and the VIF calculated for the new model.
- This process is continued till no feature has a VIF greater than 5.
- The p-values of the remaining features are observed, and any variable with a p-value greater than 0.05 (the assumed significance level) is dropped.
- In our case, none of the features had p-value greater than 0.05 after eliminating the features with VIF greater than 5.
- The desired model was obtained at the fifth iteration.

Model Evaluation

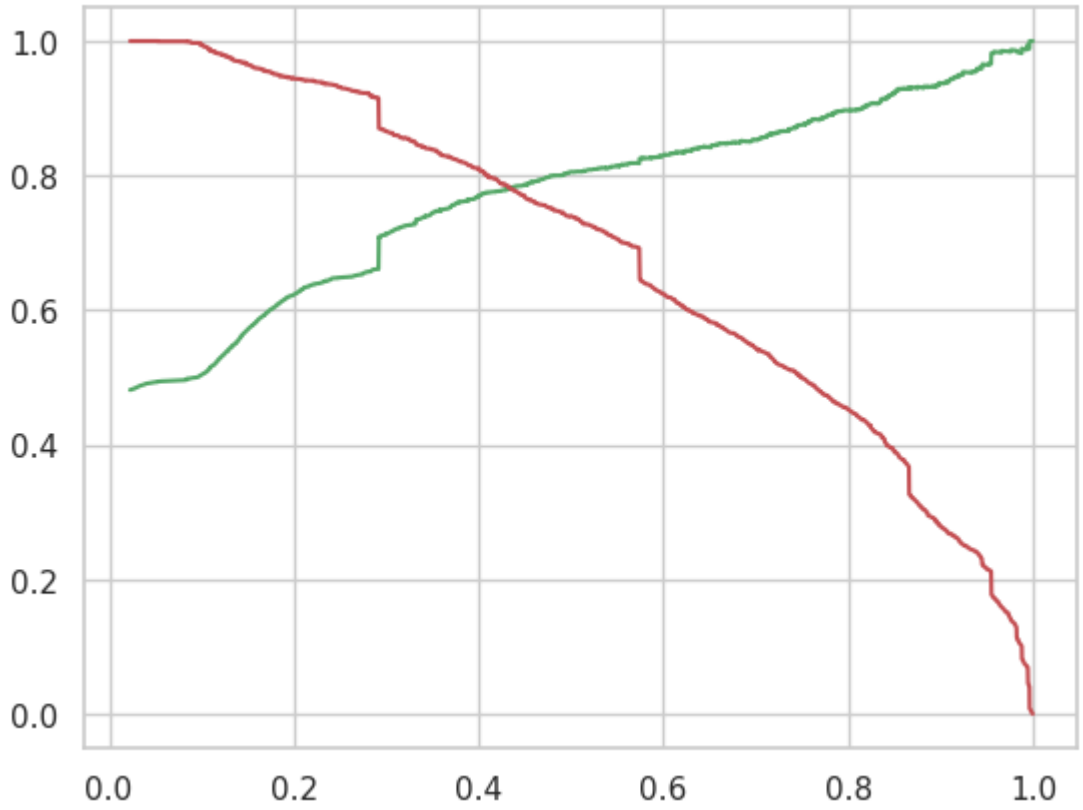
- The confusion matrix is created to evaluate how well the train set identified the correct labels for those who converted, and those who did not. Majority of the cases were correctly labelled.
- The metrics of model evaluation, such as accuracy, precision, recall, sensitivity, and specificity were calculated. Each value was found to be close to 0.8, indicating a good fitted model.
- The Receiver Operating Characteristic (ROC) Curve was constructed to display a tradeoff between sensitivity and specificity. The curve being along the left-side vertical axis and top horizontal axis, shows high accuracy.
- To obtain an optimum pair of precision and recall values, the precision-recall-tradeoff curve was obtained, which suggested 0.44 to be the suitable probability cutoff value.
- The values of each evaluation metric on the train set and test set were very close, indicating that the model was neither overfitted, nor underfitted. The model obtained is acceptable.

Precision-Recall Tradeoff, Final Train Set

Final ROC Curve

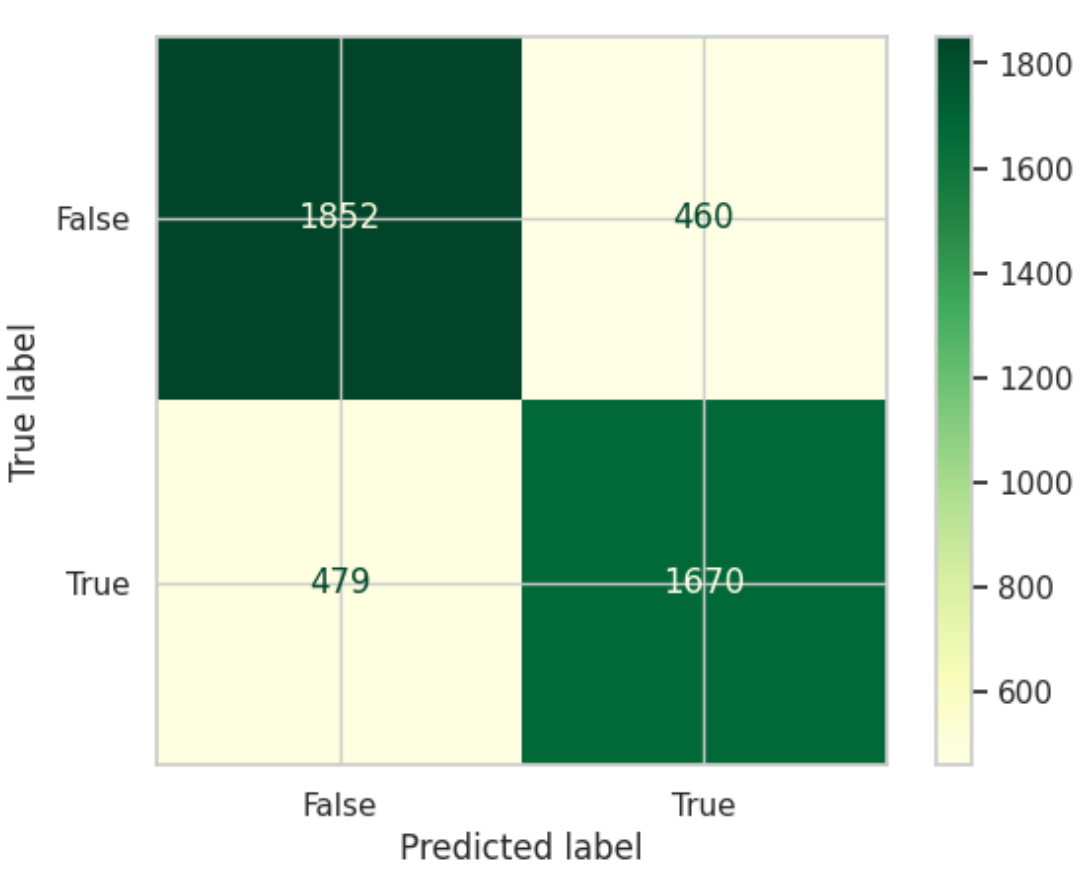


Precision-Recall Tradeoff

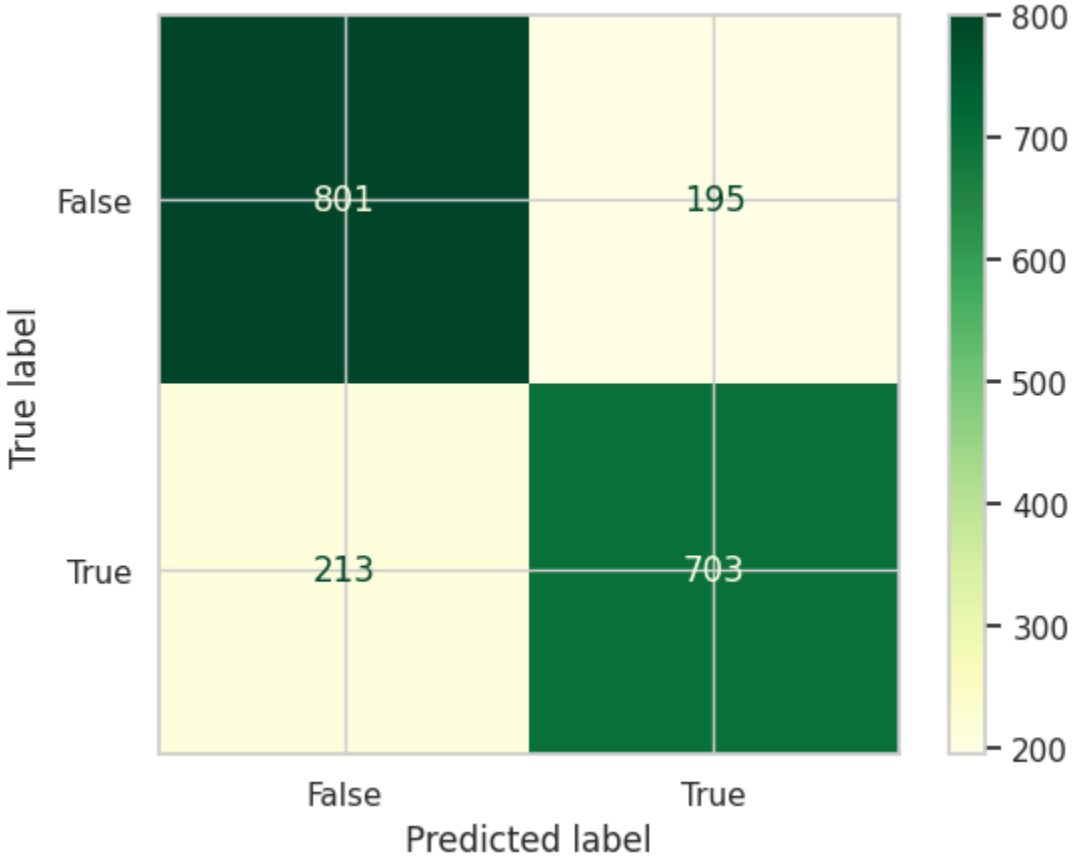


Confusion Matrix: Train Set, Test Set

Confusion Matrix: Train Set



Confusion Matrix: Test Set



Final Model and Evaluation Metrics

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	4461
Model:	GLM	Df Residuals:	4449
Model Family:	Binomial	Df Model:	11
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2079.1
Date:	Tue, 21 Nov 2023	Deviance:	4158.1
Time:	15:04:23	Pearson chi2:	4.80e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.3642
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2040	0.196	1.043	0.297	-0.179	0.587
TotalVisits	11.1489	2.665	4.184	0.000	5.926	16.371
Total Time Spent on Website	4.4223	0.185	23.899	0.000	4.060	4.785
Lead Origin_Lead Add Form	4.2051	0.258	16.275	0.000	3.699	4.712
Lead Source_Olark Chat	1.4526	0.122	11.934	0.000	1.214	1.691
Lead Source_Welingak Website	2.1526	1.037	2.076	0.038	0.121	4.185
Do Not Email_Yes	-1.5037	0.193	-7.774	0.000	-1.883	-1.125
Last Activity_Had a Phone Conversation	2.7552	0.802	3.438	0.001	1.184	4.326
Last Activity_SMS Sent	1.1856	0.082	14.421	0.000	1.024	1.347
What is your current occupation_Student	-2.3578	0.281	-8.392	0.000	-2.908	-1.807
What is your current occupation_Unemployed	-2.5445	0.186	-13.699	0.000	-2.908	-2.180
Last Notable Activity_Unreachable	2.7846	0.807	3.449	0.001	1.202	4.367

**FINAL TRAIN DATASET OUTPUT **

1. Accuracy = 0.7845188284518828
2. Precision = 0.784037558685446
3. Recall = 0.7771056305258259
4. Sensitivity = 0.7771056305258259
5. Specificity = 0.801038062283737

**FINAL TEST DATASET OUTPUT **

1. Accuracy = 0.7866108786610879
2. Precision = 0.7828507795100222
3. Recall = 0.767467248908297
4. Sensitivity = 0.767467248908297
5. Specificity = 0.8042168674698795

Recommendations

- It may be recommended that the company segments their courses to make them more affordable to students.
- Introductory and certification courses may be rolled out by the company, which would be affordable, and will also have the potential to increase the proportion of potential “Hot Leads”.
- Social media advertising may be used by the company more effectively to popularize its courses among the youth.