LEAD SCORING CASE STUDY

- Rakesh Melangi
- Syed Hanzala
- Kumar Abhirup

CONTEXT

Summary

X Education is an online education company that sells courses to industry professionals. It acquires leads through marketing efforts and referrals, and the typical lead conversion rate is 30%. The company aims to increase the lead conversion rate by identifying "Hot Leads." It has appointed you to build a model that assigns a lead score to each lead, so the leads with a higher score are more likely to convert to paying customers. You have been provided with a dataset of 9000 past leads with attributes such as lead source, time spent on the website, visits, and last activity. The target variable is "Converted" with a value of 1 for converted leads and 0 for non-converted leads. The CEO's target lead conversion rate is 80%. The categorical variables include a level called "Select" that needs to be handled

Goals

There are quite a few goals for this case study:

1. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

2. There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step.

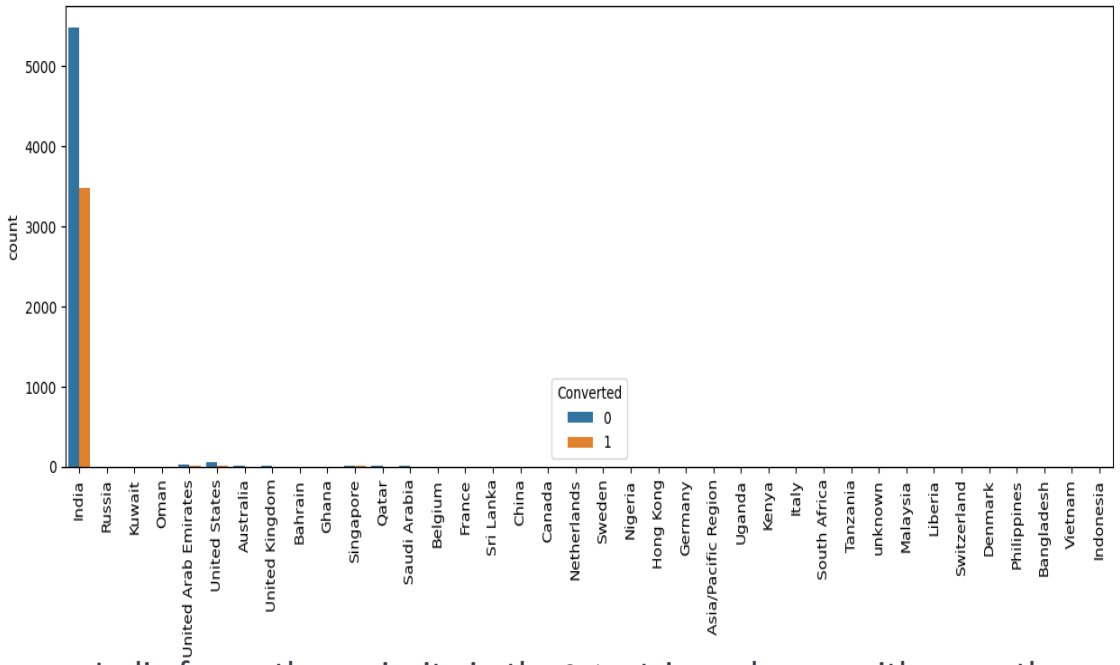
APPROACH

To build a logistic regression model to assign a lead score between 0 and 100 to each lead, the following steps can be taken:

- Data Preprocessing: Perform data exploration and cleaning as described in my previous answers.
- Feature Selection: Select the most relevant features that will be used to predict lead conversion. This can be done using techniques such as correlation analysis, chi-squared test, or mutual information.
- Data Splitting: Split the data into training and test sets in order to evaluate the model's performance on unseen data.
- Model Training: Train a logistic regression model using the selected features and the training data.
- Model Evaluation: Evaluate the performance of the model using metrics such as accuracy, precision, recall, and F1-score.
- Hyperparameter tuning: Adjust the parameters of the model to optimize its performance.
- Lead Scoring: Assign a lead score to each lead using the trained model. The logistic regression model will output a probability of conversion for each lead, which can be mapped to a score between 0 and 100.

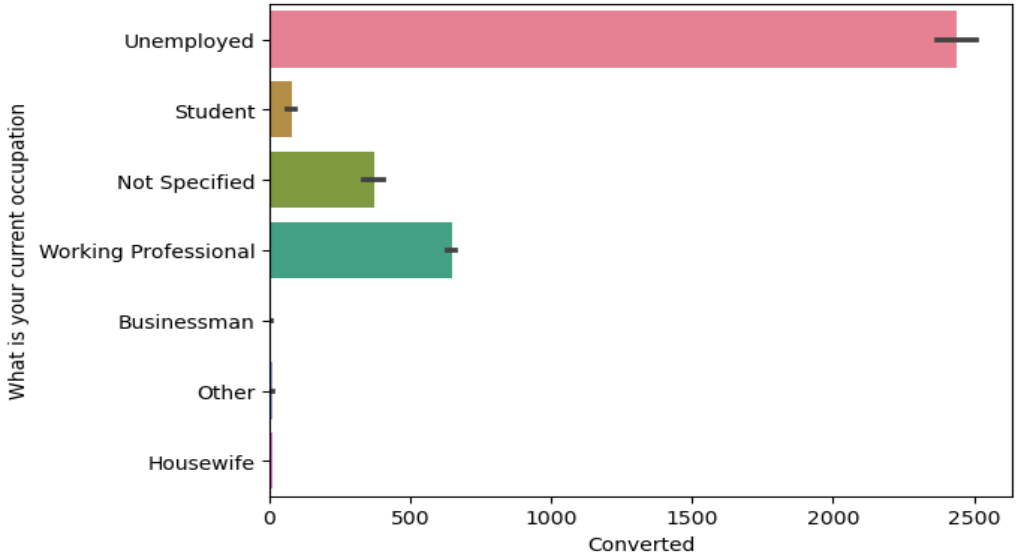
IMPUTING MISSING VALUES AND DROPPING UNBALANCED COLUMNS

COUNTRY



India forms the majority in the countries columns with more than 97% of the data. Hence there is now value being added by this column to the over all analysis. Hence dropping this column

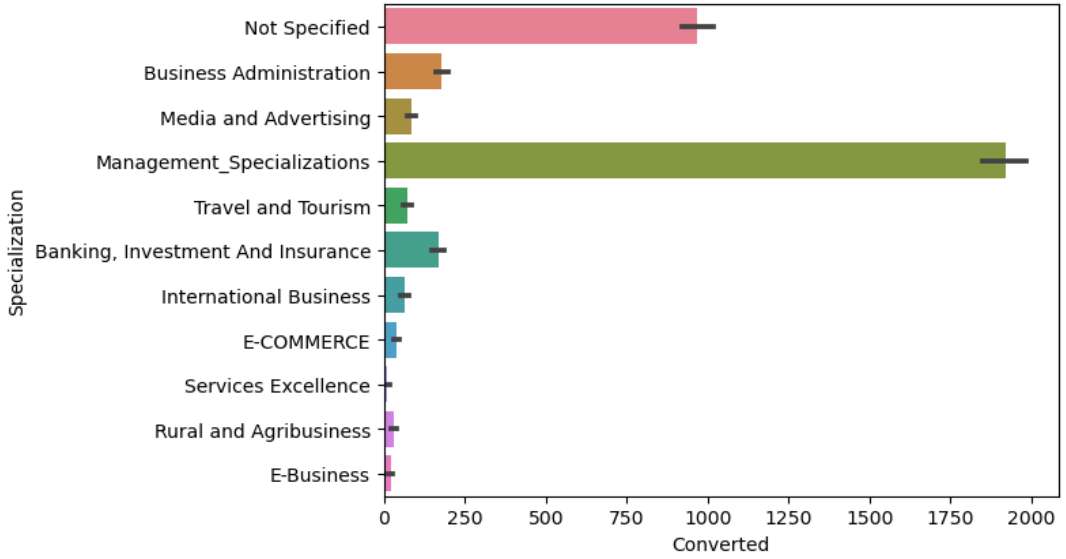
SPECIALIZATION



Since all the management specializations show similar trend we can classify them under Management Specialization. This reduces the complexity of the model. Since there is an option to not specify the option we fill in Non Specified for NaN

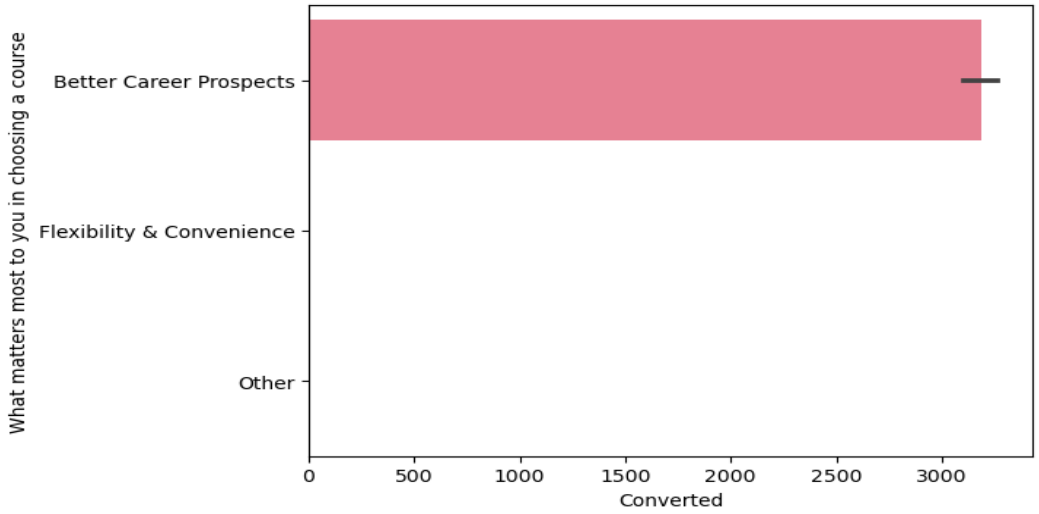
IMPUTING MISSING VALUES AND DROPPING UNBALANCED COLUMNS

What is your current occupation'



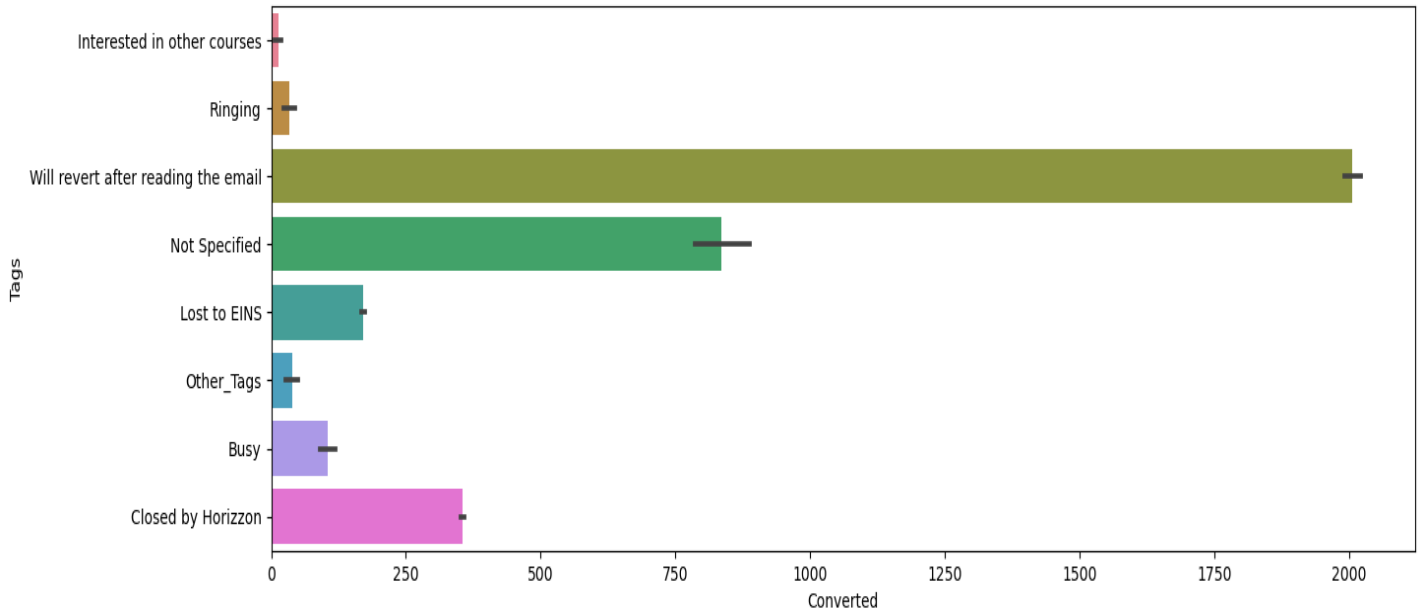
Filling in Not Specified for the Null values since occupation can have an impact on the upskilling. This becomes important. Hence the not specified for NaN

What matters most to you in choosing a course



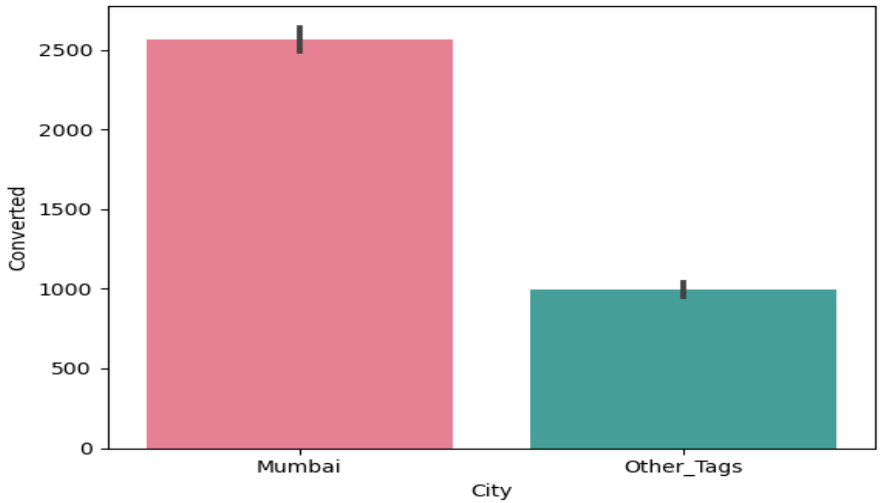
Majority of the comments seems to be towards better career prospects. But this seems to be tilted towards a single answer hence this can be dropped

IMPUTING MISSING VALUES AND DROPPING UNBALANCED COLUMNS



Tags

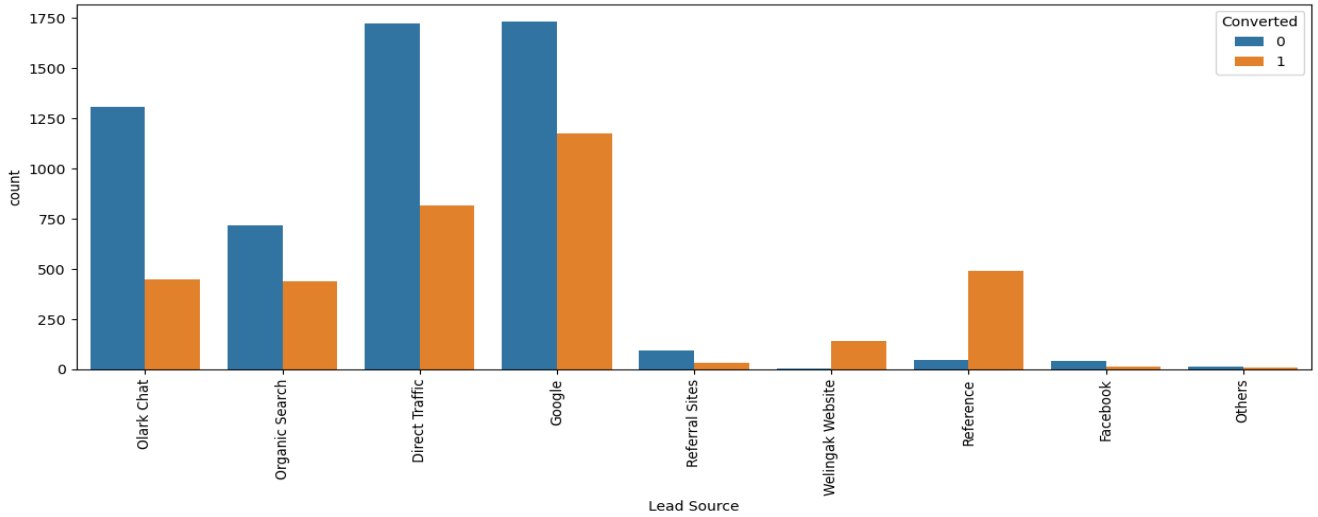
Tags could be important because they are inputs from the employees and can be based on their intuition. Hence not dropping the same. For the smaller distributions they can be classified under Other_tags



City

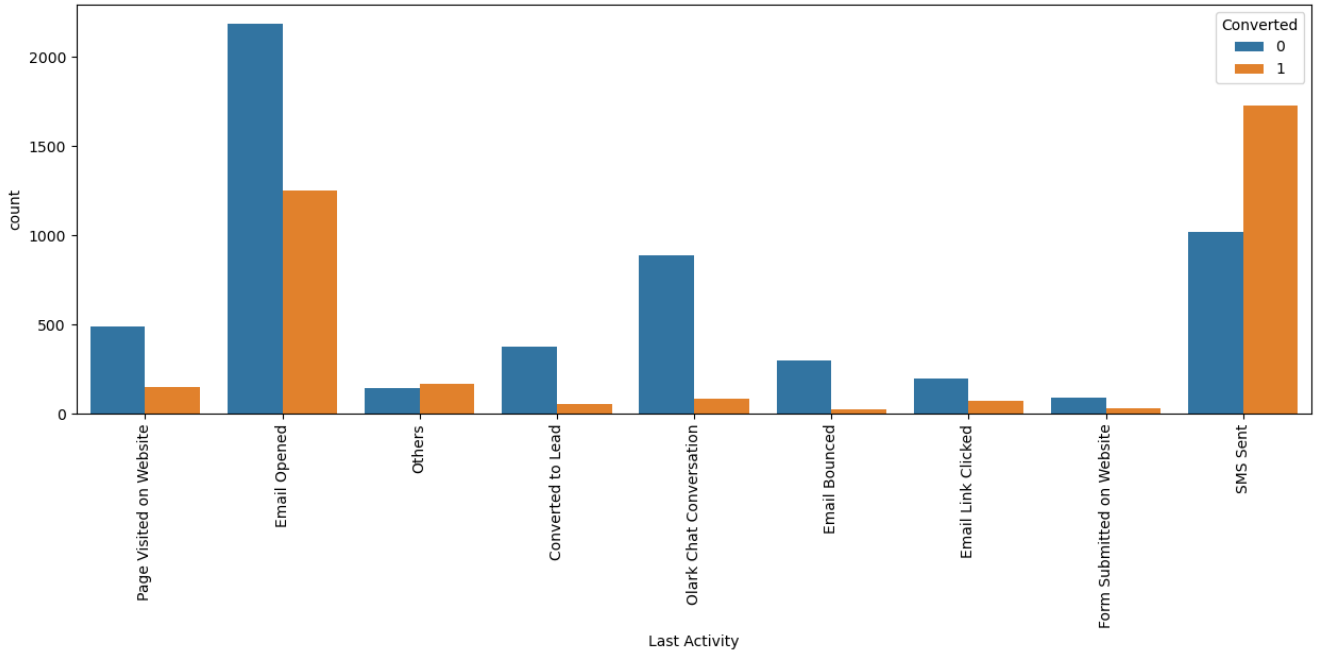
Mumbai has the majority in terms of the city. Hence using it to fill in the blanks and to reduce complexity we can tag it to Other_cities

IMPUTING MISSING VALUES AND DROPPING UNBALANCED COLUMNS



Lead Source

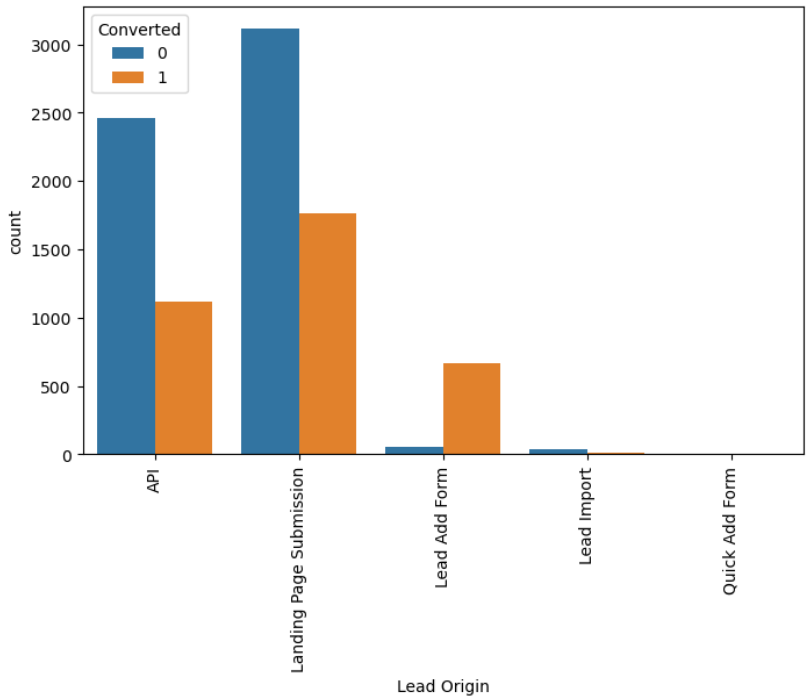
Since Google is the largest bucket we can impute that to replace NaN. We can categorize other leads which are smaller to Other



Last Activity

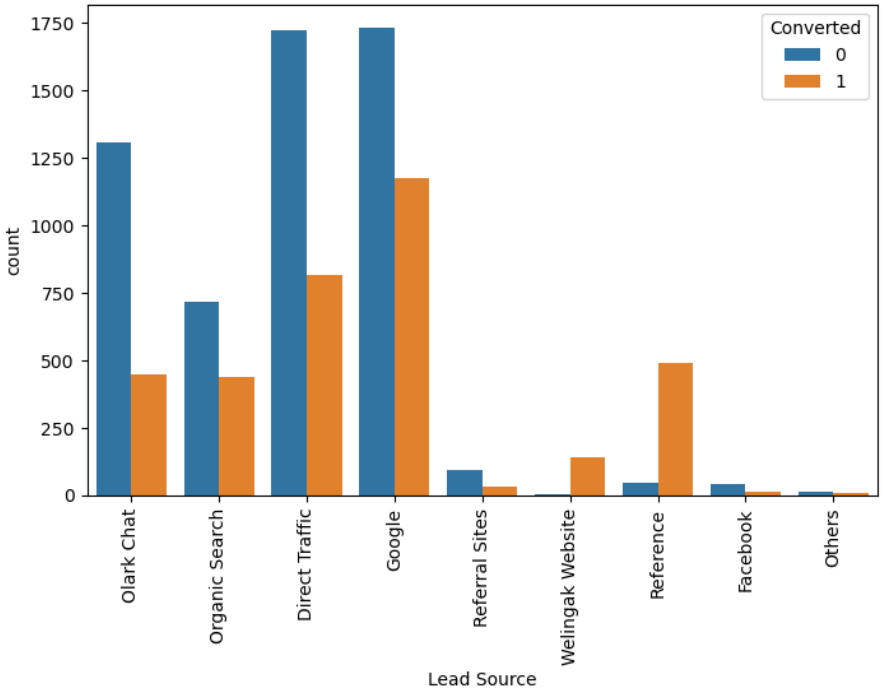
Replacing using the median since this is well distributed across multiple areas.

UNDERSTANDING OTHER DATA COLUMNS



Last Origin

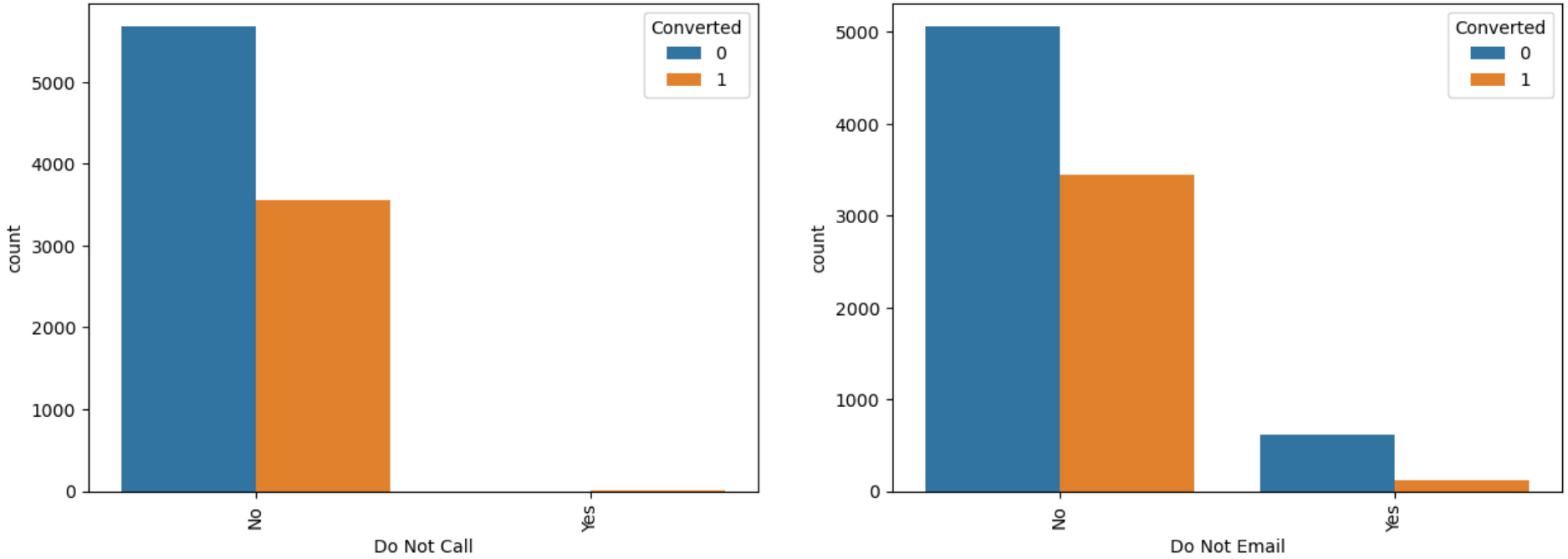
Inference API and Landing Page Submission bring higher number of leads as well as conversion. Lead Add Form has a very high conversion rate but count of leads are not very high. Lead Import and Quick Add Form get very few leads.



Last Source

Google and Direct seems to be higher in terms of both traffic and conversions. Organic search has the highest probability of conversion whereas Facebook and Referrals seems to be the lowest

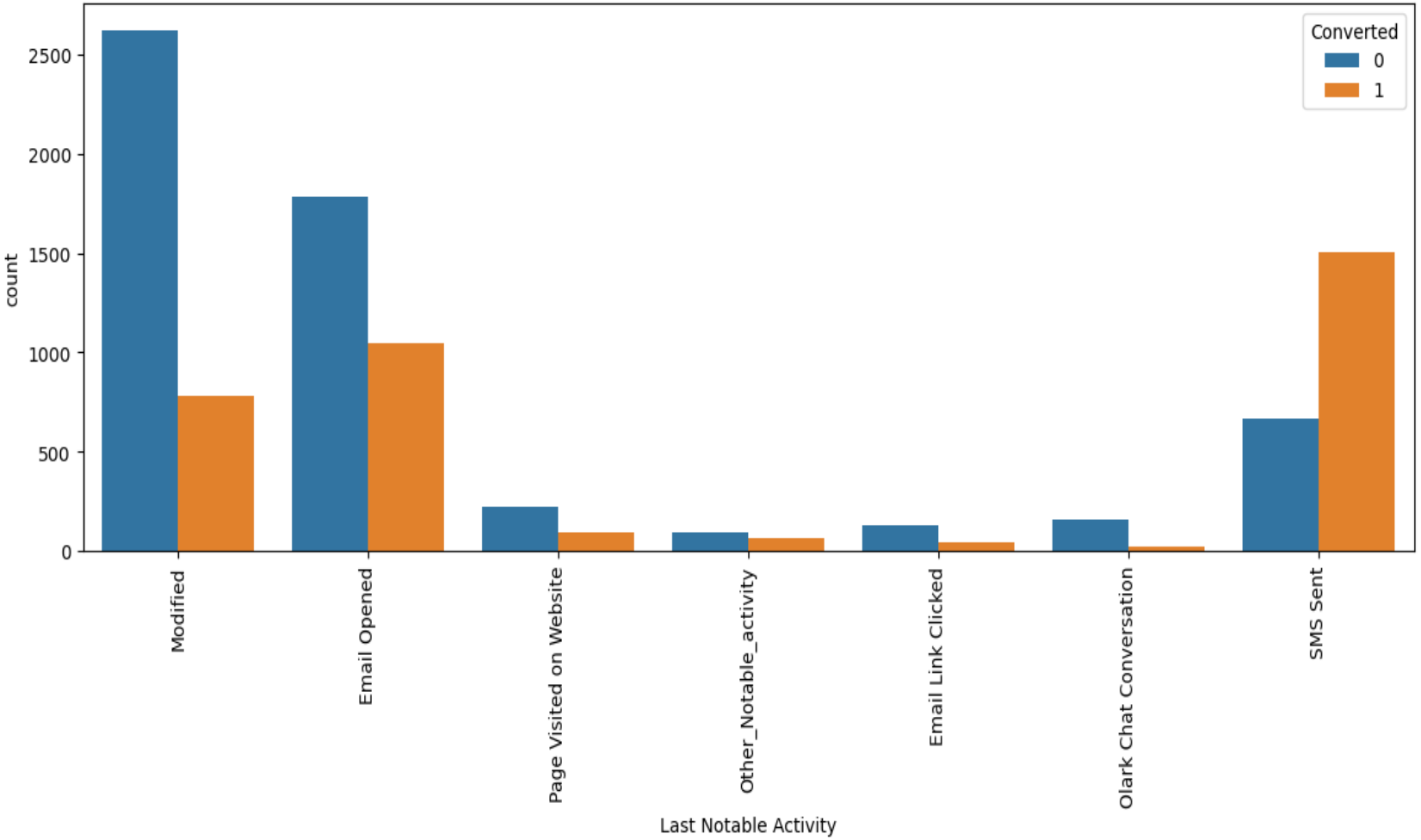
UNDERSTANDING OTHER DATA COLUMNS



Do Not Call & Do Not Email

Preferences of not calling and not emailing seems to not have much impact. On the contrary it seems to have higher traffic and conversions as compared to the alternative

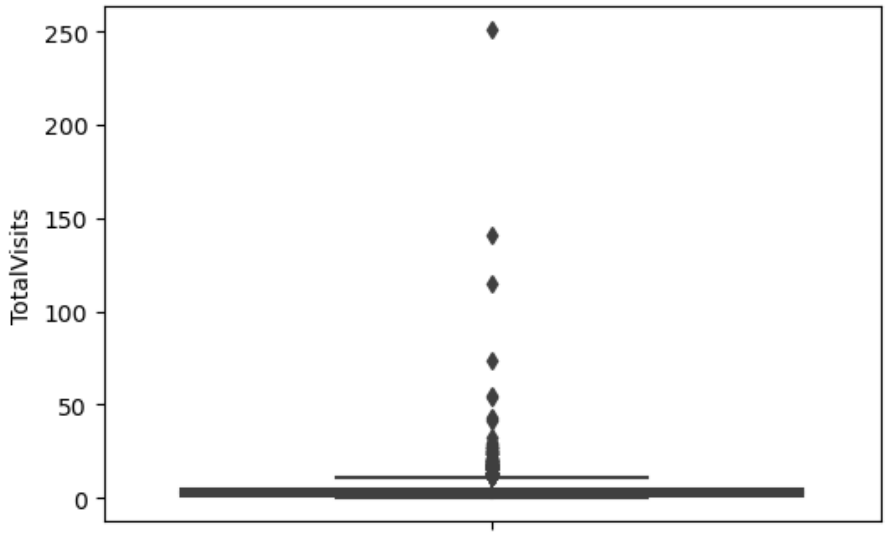
UNDERSTANDING OTHER DATA COLUMNS



Last Notable Activity

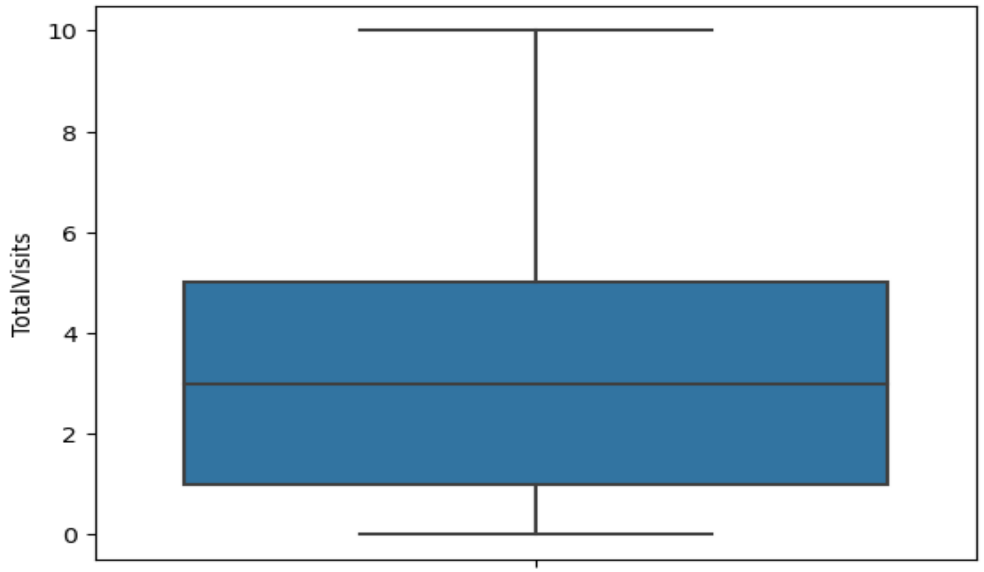
SMS sent seems to be have the highest conversion rate as compared to Others. Modified is highest in rejections and next come email. So combination of these three might help in higher and targeted conversions.

OUTLIER TREATMENT



Summary Statistics

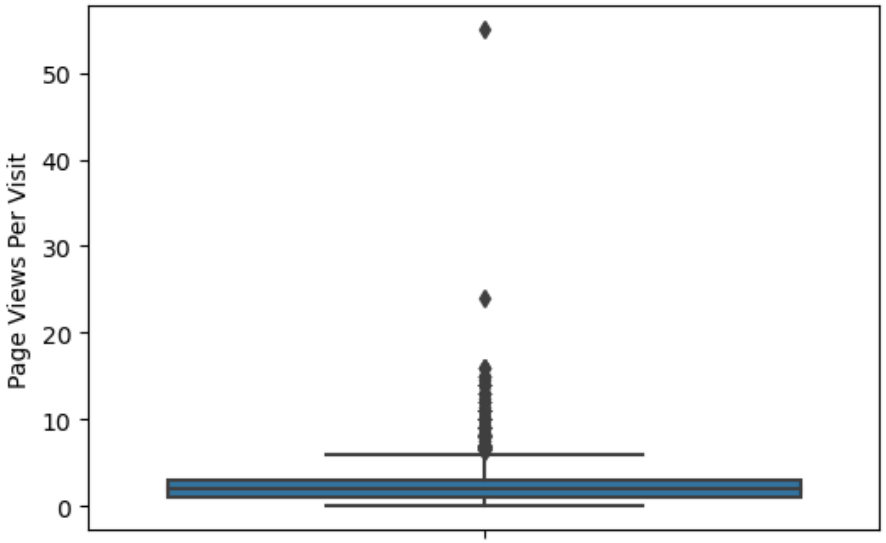
count	9240.000000
mean	3.438636
std	4.819024
min	0.000000
5%	0.000000
25%	1.000000
50%	3.000000
75%	5.000000
90%	7.000000
95%	10.000000
99%	17.000000
max	251.000000



Total Visits

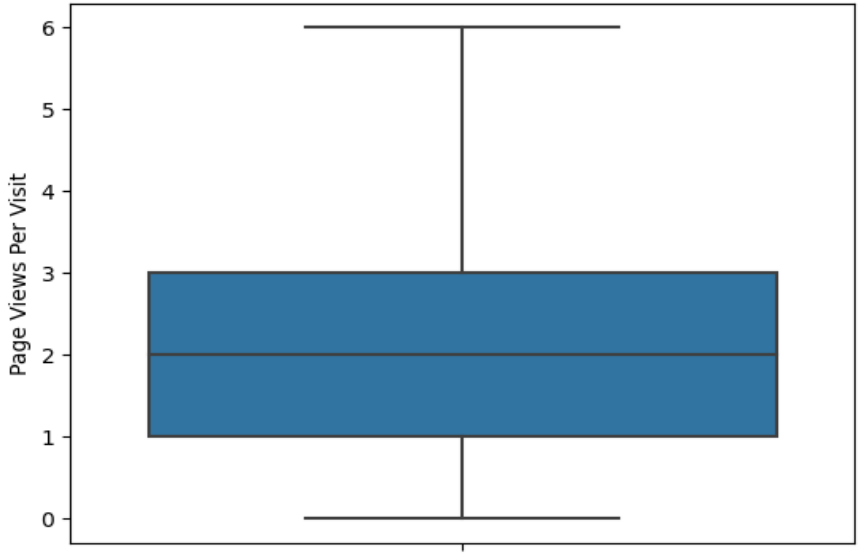
Taking percentiles and ensuring we cap the values to 5% and 95% to address the outliers

OUTLIER TREATMENT



Summary Statistics

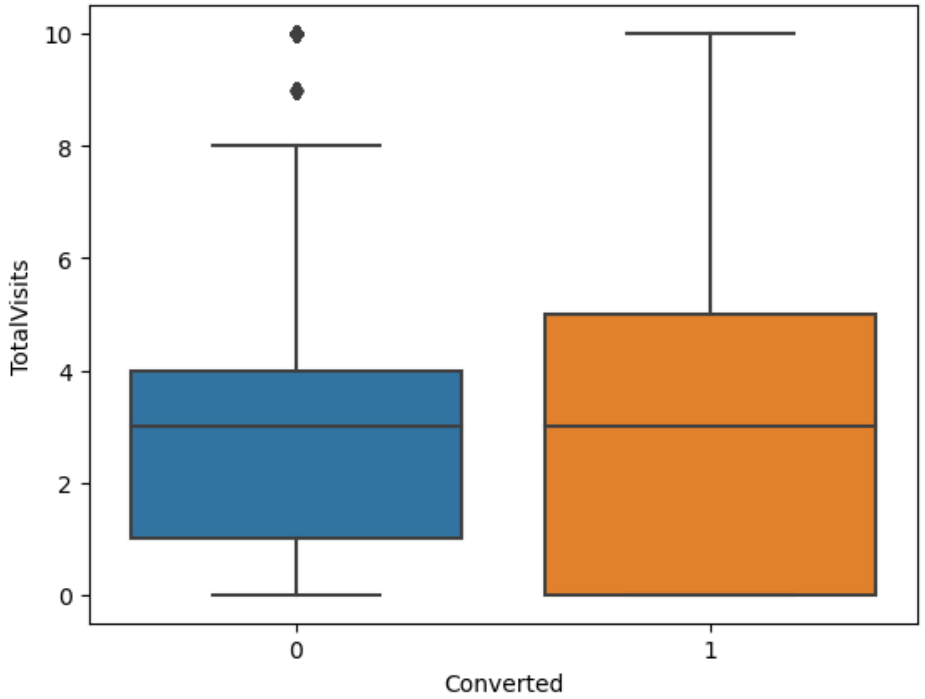
count	9240.000000
mean	2.357440
std	2.145781
min	0.000000
25%	1.000000
50%	2.000000
75%	3.000000
max	55.000000



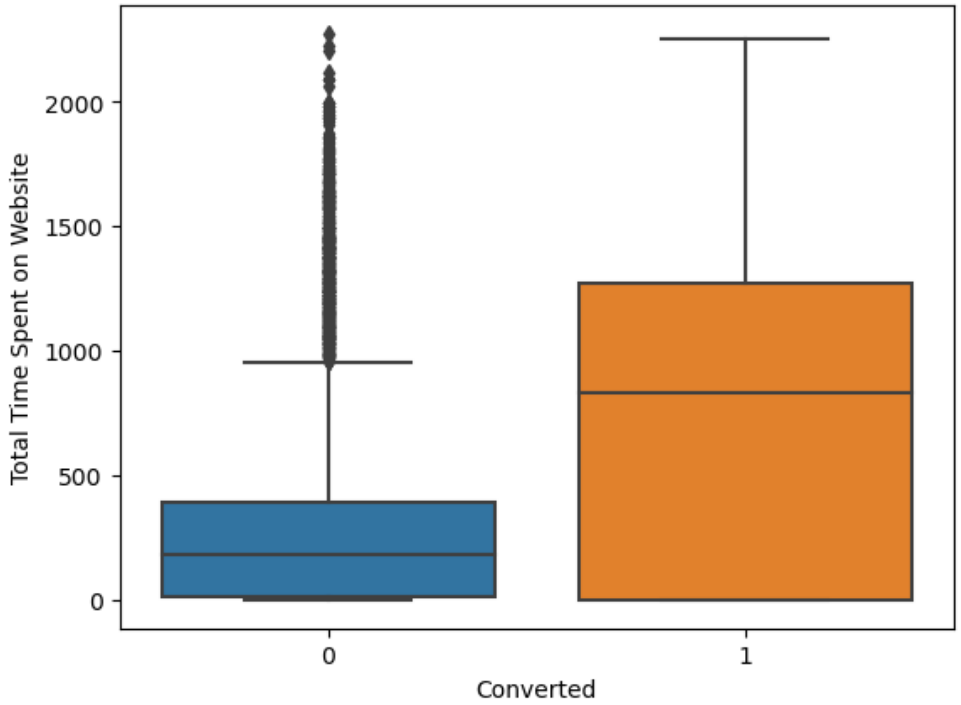
Total Visits

Taking percentiles and ensuring we cap the values to 5% and 95% to address the outliers

OUTLIER TREATMENT



Inference Median for converted and not converted leads are the close.



Inference Website should be made more engaging as Leads spending more time on the website are more likely to be converted, so to make leads spend more time

BOTH P AND VIF VALUES ARE LOW AND HENCE GOOD TO PROCEED

Linear Regression & VIF Evaluation

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6449
Model Family:	Binomial	Df Model:	18
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1142.0
Date:	Tue, 31 Jan 2023	Deviance:	2284.0
Time:	21:09:04	Pearson chi2:	1.10e+04
No. Iterations:	8	Pseudo R-squ. (CS):	0.6233
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2737	0.149	1.838	0.066	-0.018	0.566
Total Time Spent on Website	1.0221	0.062	16.395	0.000	0.900	1.144
Lead Origin_Lead Add Form	1.0710	0.357	2.996	0.003	0.370	1.772
What is your current occupation_Not Specified	-2.5330	0.157	-16.083	0.000	-2.842	-2.224
Specialization_Travel and Tourism	-1.0630	0.448	-2.374	0.018	-1.941	-0.185
Lead Source_Olark Chat	0.9284	0.156	5.945	0.000	0.622	1.234
Lead Source_Welingak Website	2.5718	0.820	3.138	0.002	0.965	4.178
Tags_Busy	-1.4662	0.265	-5.542	0.000	-1.985	-0.948
Tags_Closed by Horizzon	5.6333	1.031	5.464	0.000	3.613	7.654
Tags_Interested in other courses	-3.4708	0.368	-9.426	0.000	-4.193	-2.749
Tags_Lost to EINS	5.7445	0.755	7.608	0.000	4.265	7.224
Tags_Other_Tags	-4.2260	0.241	-17.500	0.000	-4.699	-3.753
Tags_Ringing	-5.2752	0.270	-19.534	0.000	-5.804	-4.746
Tags_Will revert after reading the email	2.7002	0.214	12.616	0.000	2.281	3.120
Last Activity_Email Bounced	-1.2627	0.411	-3.073	0.002	-2.068	-0.457
Last Activity_SMS Sent	2.0202	0.127	15.926	0.000	1.772	2.269
Last Notable Activity_Email Link Clicked	-1.1805	0.470	-2.511	0.012	-2.102	-0.259
Last Notable Activity_Modified	-1.4857	0.130	-11.445	0.000	-1.740	-1.231
Last Notable Activity_Olark Chat Conversation	-1.5750	0.483	-3.398	0.001	-2.483	-0.667

	Features	VIF
16	Last Notable Activity_Modified	1.96
1	Lead Origin_Lead Add Form	1.80
2	What is your current occupation_Not Specified	1.64
12	Tags_Will revert after reading the email	1.63
14	Last Activity_SMS Sent	1.59
4	Lead Source_Olark Chat	1.53
0	Total Time Spent on Website	1.40
7	Tags_Closed by Horizzon	1.33
5	Lead Source_Welingak Website	1.26
10	Tags_Other_Tags	1.26
8	Tags_Interested in other courses	1.15
11	Tags_Ringing	1.14
13	Last Activity_Email Bounced	1.10
17	Last Notable Activity_Olark Chat Conversation	1.08
6	Tags_Busy	1.05
15	Last Notable Activity_Email Link Clicked	1.05
9	Tags_Lost to EINS	1.05
3	Specialization_Travel and Tourism	1.03

- P and VIF (Variance Inflation Factor) are both used in Logistic Regression to check for multicollinearity.
- P-value indicates the significance of each predictor in the model. A low P-value (< 0.05) suggests that the predictor is significant, while a high P-value (> 0.05) suggests that the predictor is not significant.
- VIF measures the multicollinearity between predictors. High VIF values (typically > 5) indicate that a predictor is highly correlated with other predictors in the model, which can affect the interpretation and reliability of the model. Checking both P and VIF values helps in selecting a subset of predictors that are both significant and uncorrelated with other predictors.

CONFUSION MATRIX AND INITIAL VALUES

Predicted / Actual	Not Converted	Converted
Not Converted	3844	158
Converted	255	2211

The model seems to be performing well based on the above table

Overall Accuracy : 93.61%
Sensitivity: 89.6%
Specificity: 96.1%
False Positive Rate:3.9%
+ve Prdictive Value: 93.3%
-Ve Predictive Rate: 93.7%

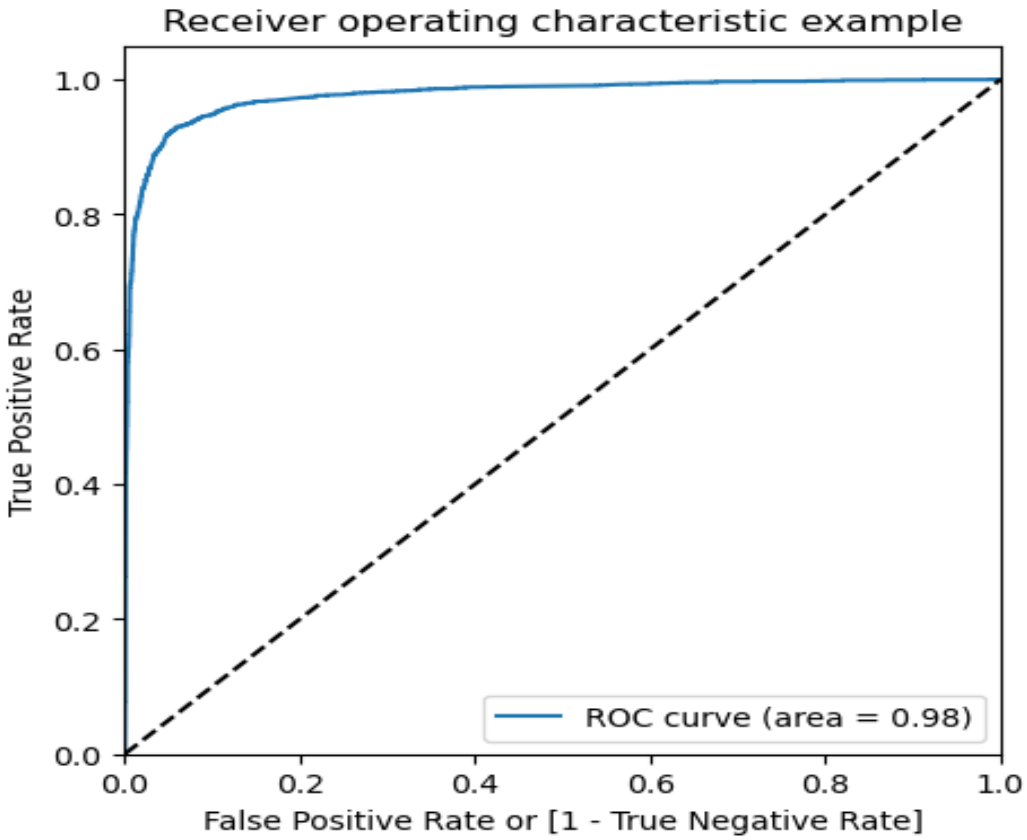
- A confusion matrix is a table used to evaluate the performance of a classification algorithm. In logistic regression, a confusion matrix is created by comparing the predicted class values to the actual class values for a set of test data. It has 4 main components:
- True Positives (TP)
- Number of instances where the actual class is positive and the predicted class is also positive.

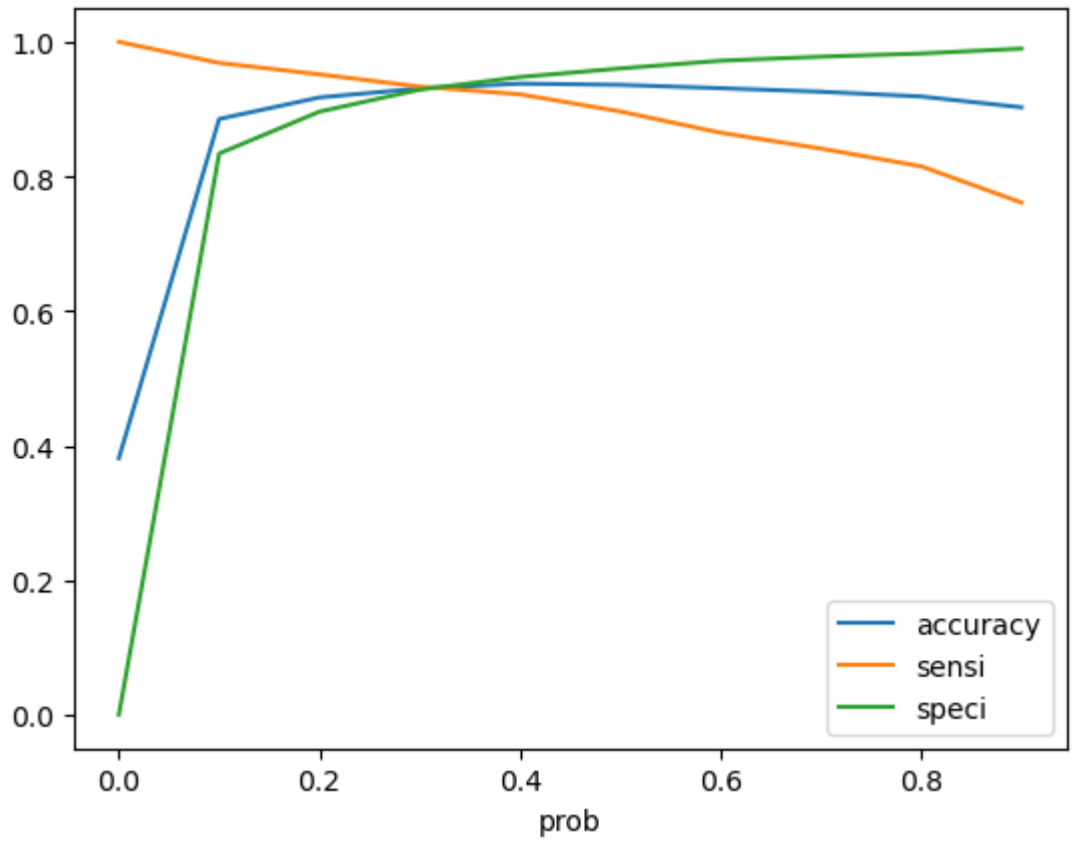
- False Positives (FP)
- Number of instances where the actual class is negative but the predicted class is positive.

- True Negatives (TN)
- Number of instances where the actual class is negative and the predicted class is also negative.

- False Negatives (FN)
- Number of instances where the actual class is positive but the predicted class is negative.

ROC CURVE





- ROC AUC score of 0.98 means that the classifier has a 98% accuracy in distinguishing between positive and negative classes.

- From the curve above, 0.3 is the optimum point to take it as a cutoff probability.

Model Evaluation: Final Result

Train Data

Metrics	
Overall Accuracy	93%
Sensitivity	93.3%
Specificity	92.9%
Based on Confusion Matrix	
False Positive rate	7%
Positive Predicted Rate	89%
Negative Predicted rate	95.7%
Precision and Recall	
Precision	89%
Recall	93.3%

Test Data

Metrics	
Overall Accuracy	93.4%
Sensitivity	95.2%
Specificity	92.3%
Precision and Recall	
Precision	88%
Recall	95.2%

Thank you