LEAD SCORE CASE STUDY

LOGISTIC REGRESSION

Naveen Katroliya
Parth Goyal

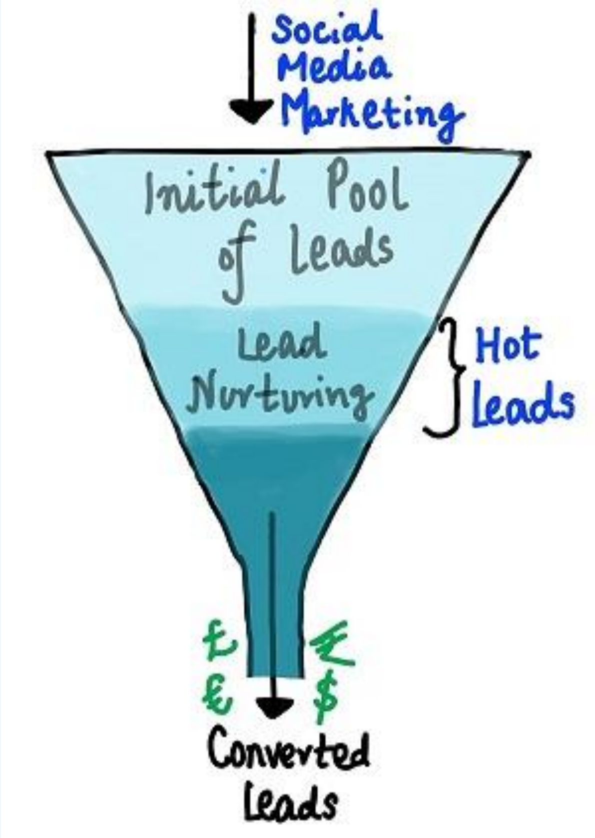
PROBLEM STATEMENT

X Education is an organization which provides online courses for industry professional. The company marks its courses on several popular websites like google.

X Education wants to select most promising leads that can be converted to paying customers.

Although the company generates a lot of leads only a few are converted into paying customers, wherein the company wants a higher lead conversion. Leads come through numerous modes like email, advertisements on websites, google searches etc.

The company has had 30% conversion rate through the whole process of turning leads into customers by approaching those leads which are to be found having interest in taking the course. The implementation process of lead generating attributes are not efficient in helping conversions.



BUSINESS GOAL

The company requires a model to be built for selecting most promising leads.

Lead score to be given to each leads such that it indicates how promising the lead could be. The higher the lead score the more promising the lead to get converted, the lower it is the lesser the chances of conversion

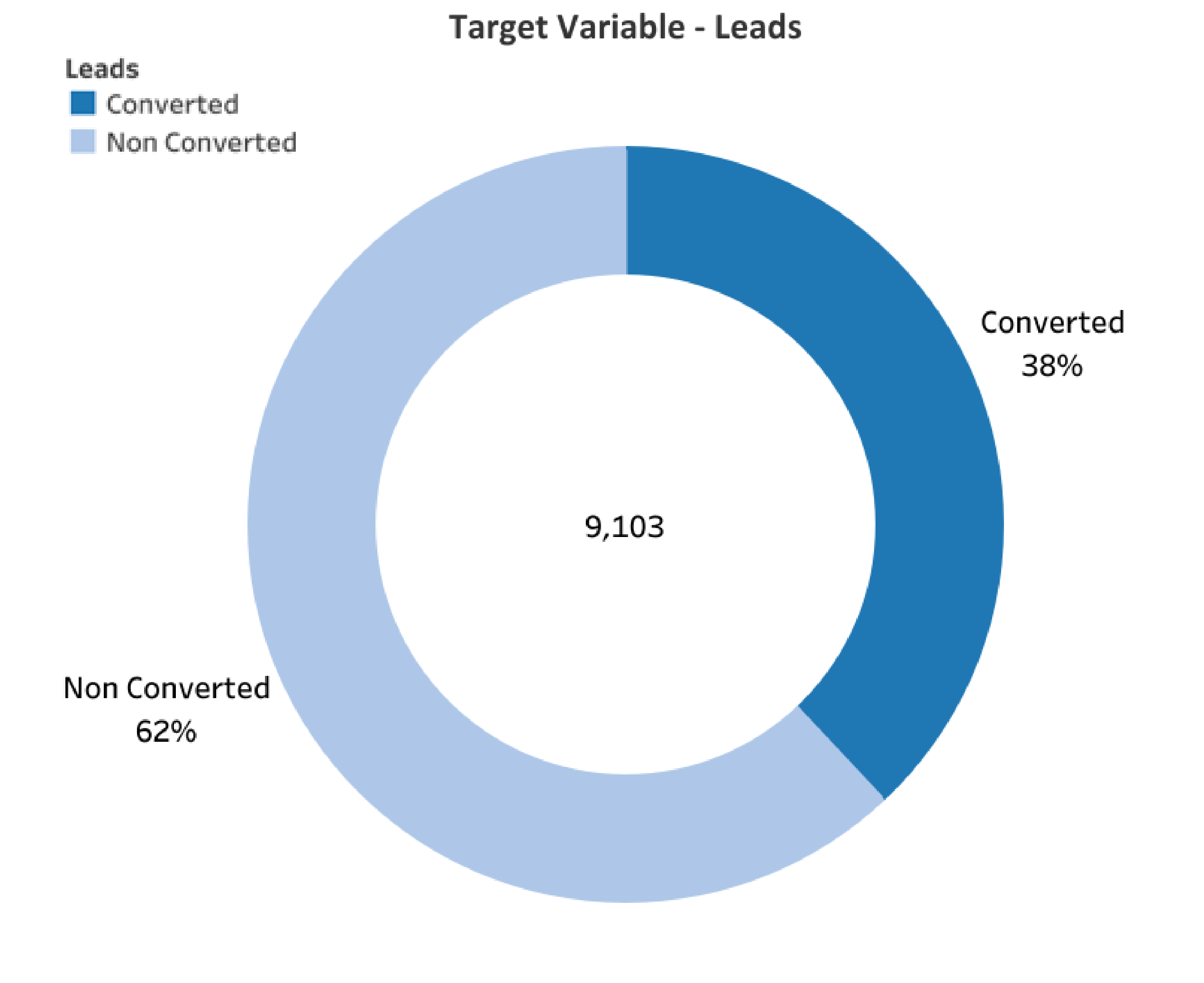
The model to be built in lead conversion rate around 80% or more.

SOLUTION APPROACH

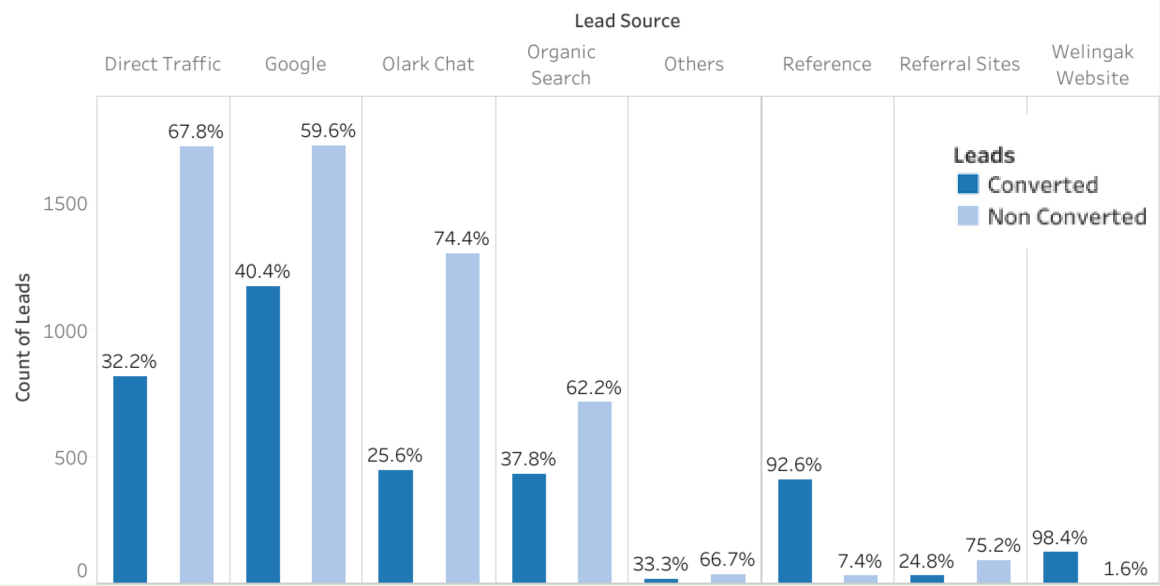
- Importing the required libraries
- Reading and understanding the data
- Data Cleaning
- Exploratory Data Analysis (EDA)
  - ✓ Univariate Analysis
  - ✓ Bivariate Analysis
- Outlier Analysis
- Data Preparation
- Splitting the Data into Training and Testing Sets
- Feature Scaling
- Model Building using Logistic Regression
- Feature selection using RFE
- Making Predictions on test sets
- Model Evaluation

TARGET VARIABLE

- We have total 9240 entries of unique clients and we need to identify out of these which have the highest probability of getting converted.
- We deleted a few entries throughout the data cleaning procedure, maintaining 9103 entries overall.
- Approximately 38% of clients were converted, as seen in the donut figure, whereas 62% were not.

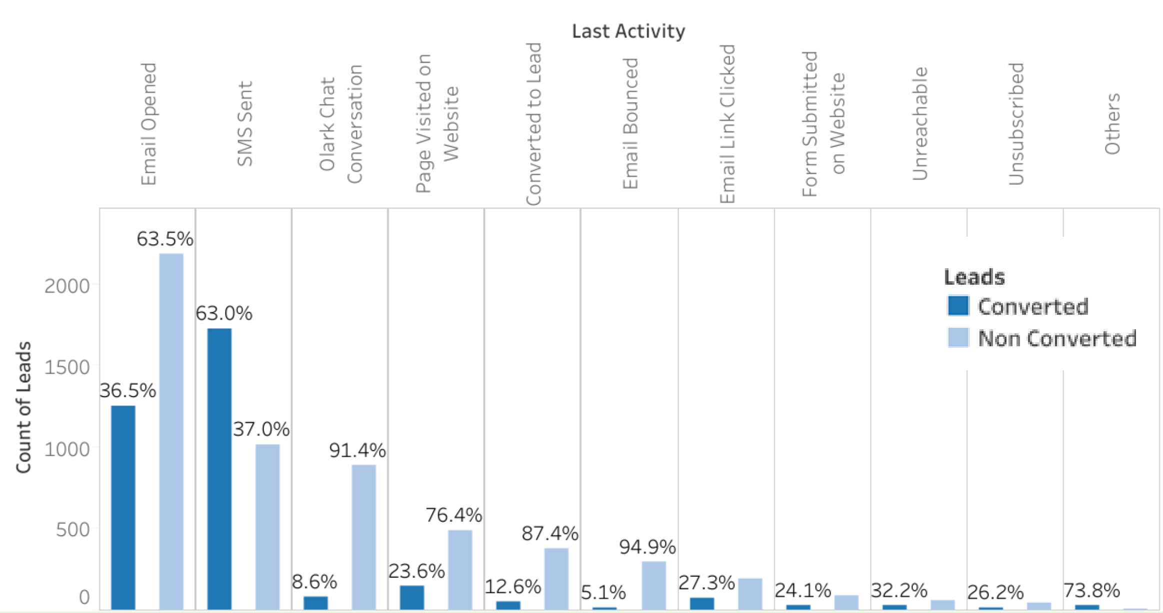


EXPLORATORY DATA ANALYSIS



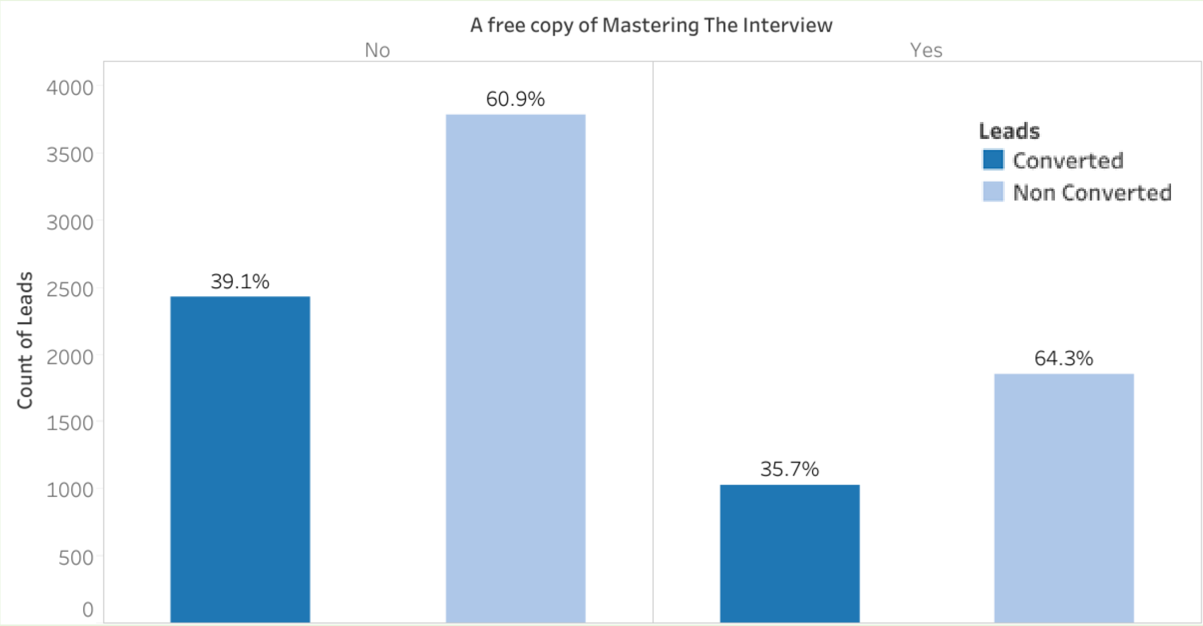
LEAD SOURCE VS CONVERTED

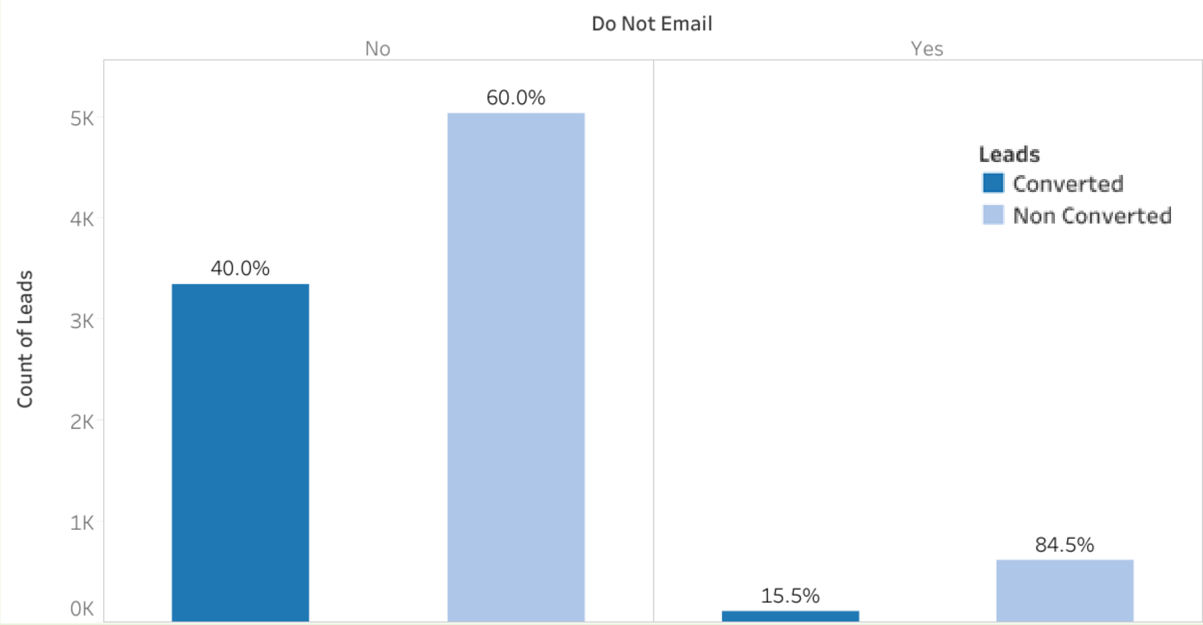
The customer arriving through the "Welingak website," "Reference," and "Google" has an extraordinarily high conversion rate.



LAST ACTIVITY VS CONVERTED

Around 63% of the clients in the "SMS sent" category have converted, and 36% of the consumers in the "Email opened" category.





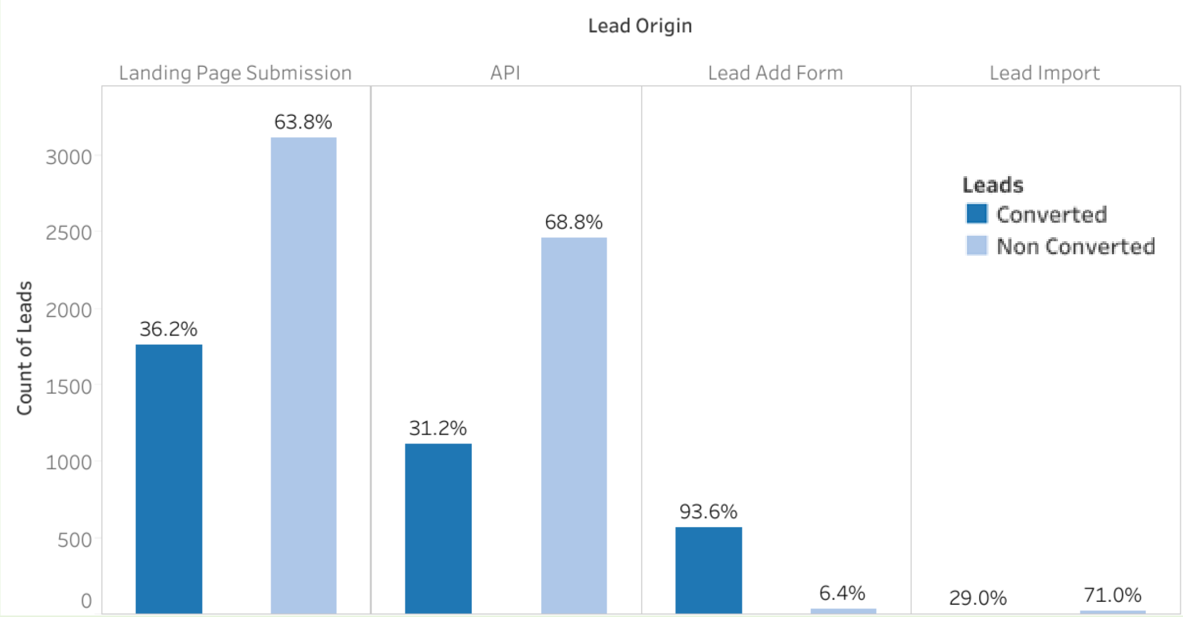
A FREE COPY OF MASTERING THE INTERVEIW VS CONVERTED

The conversion graph for "A free copy of mastering the interview" revealed no discernible trends or insights.

DO NOT EMAIL VS CONVERTED

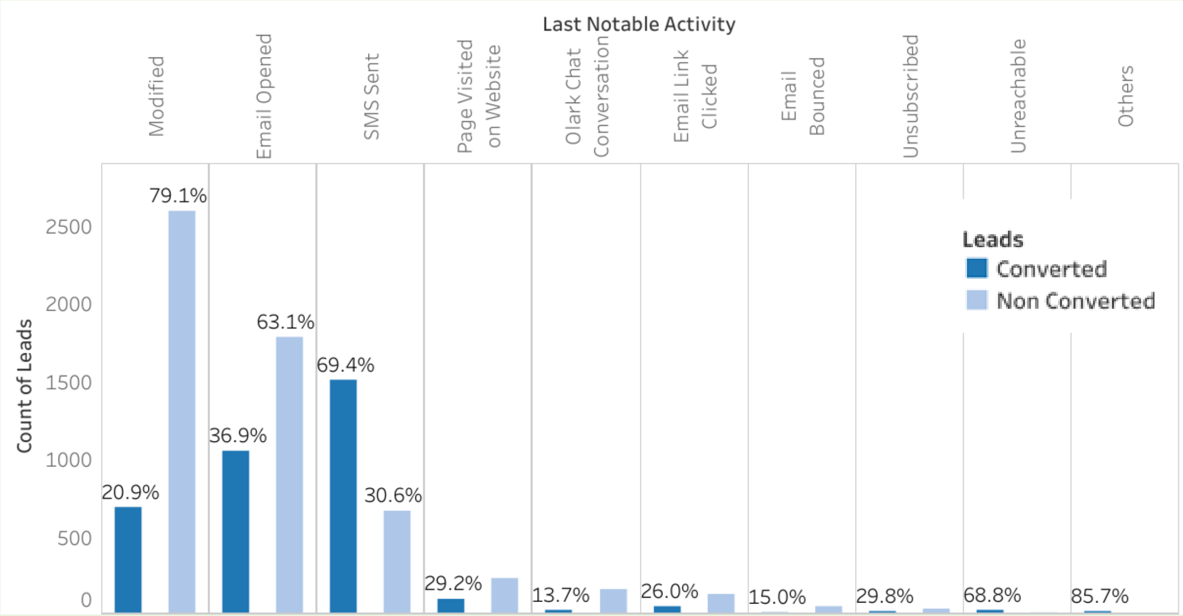
40% of customers who choose to receive emails about the course end up enrolling.

LEAD ORIGIN VS CONVERTED

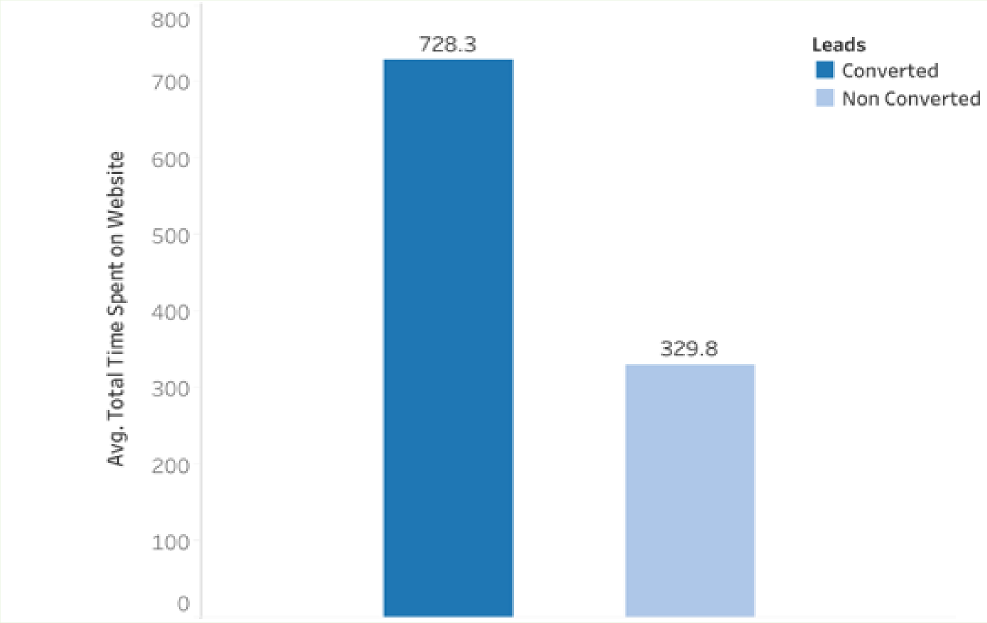


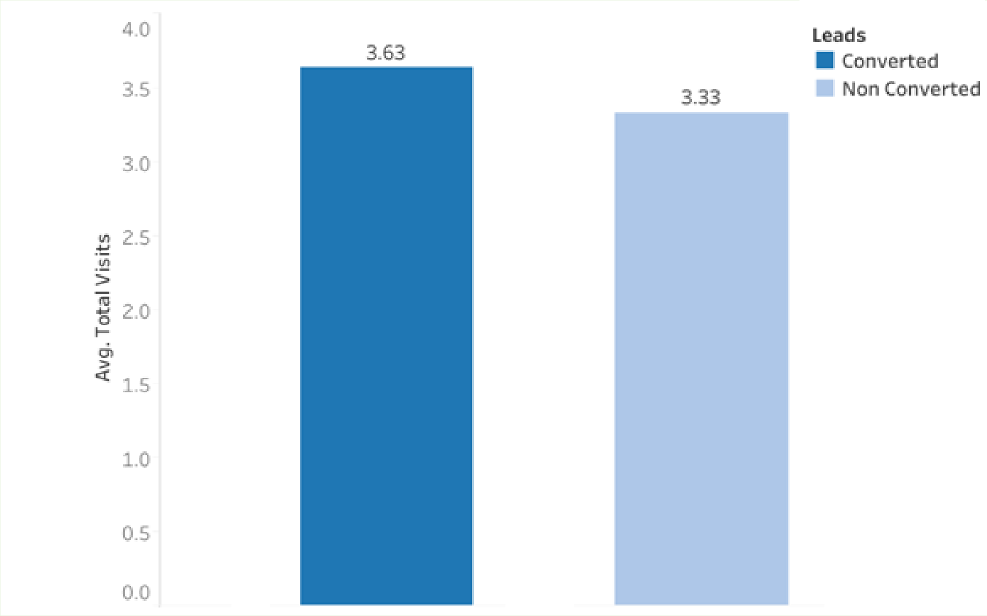
The conversion rate for the "Lead Add Form" category appears to be 93%, and 36% for the "Landing Page Submission" category.

LAST NOTABLE ACTIVITY VS CONVERTED



Approximately 69% of clients in the "SMS sent" category and 36% of clients in the "Email opened" category have converted; nevertheless, the numbers are comparable to the "Last action" conversion plot.





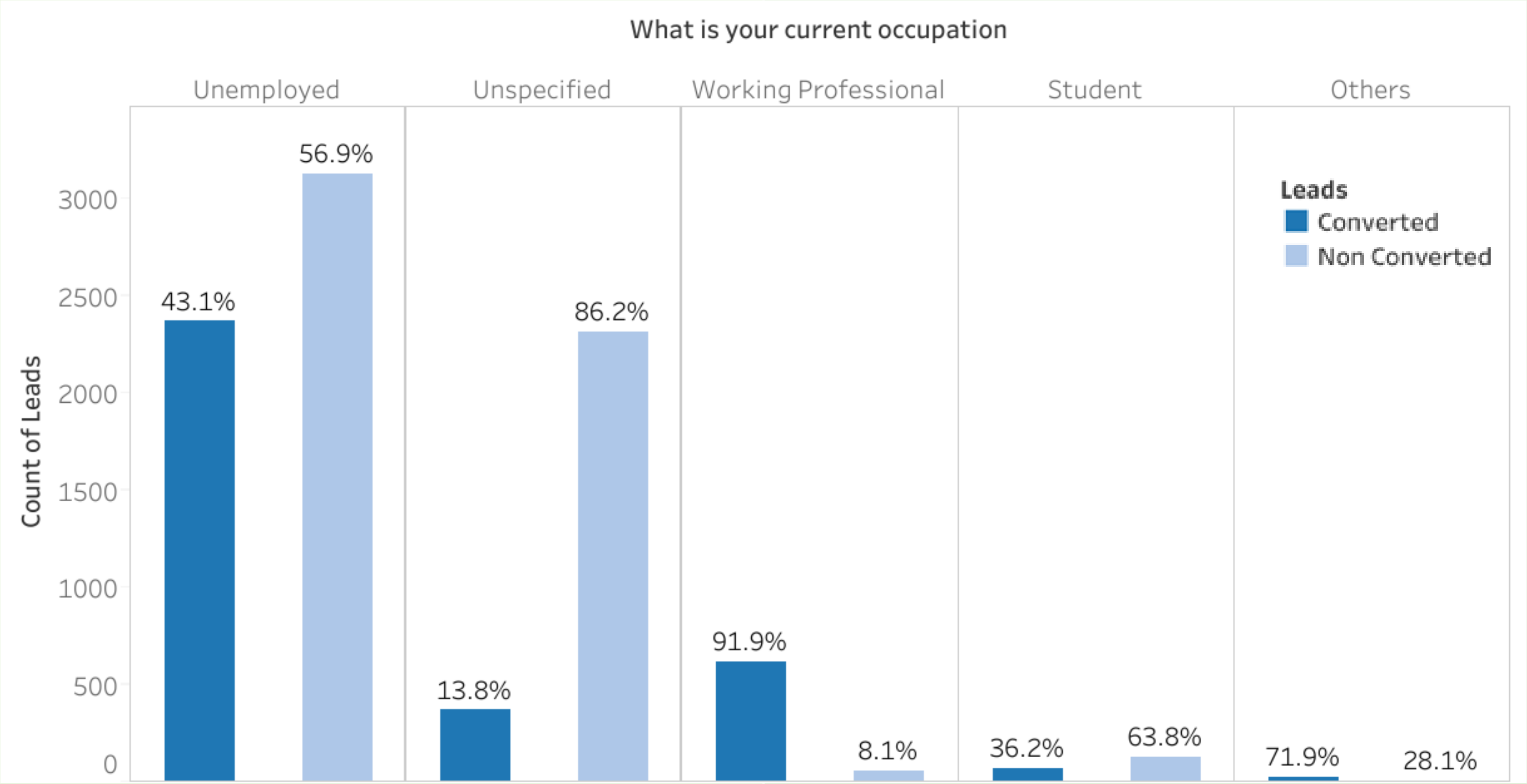
TOTAL TIME SPENT ON WEBSITES VS CONVERTED

The average "total time spent on website" for those who converted was roughly 728, compared to 328 for those who weren't converted.

TOTAL VISITS VS CONVERTED

There isn't a huge difference between converted and non-converted ones, as the data indicates.

OCCUPATION VS CONVERTED

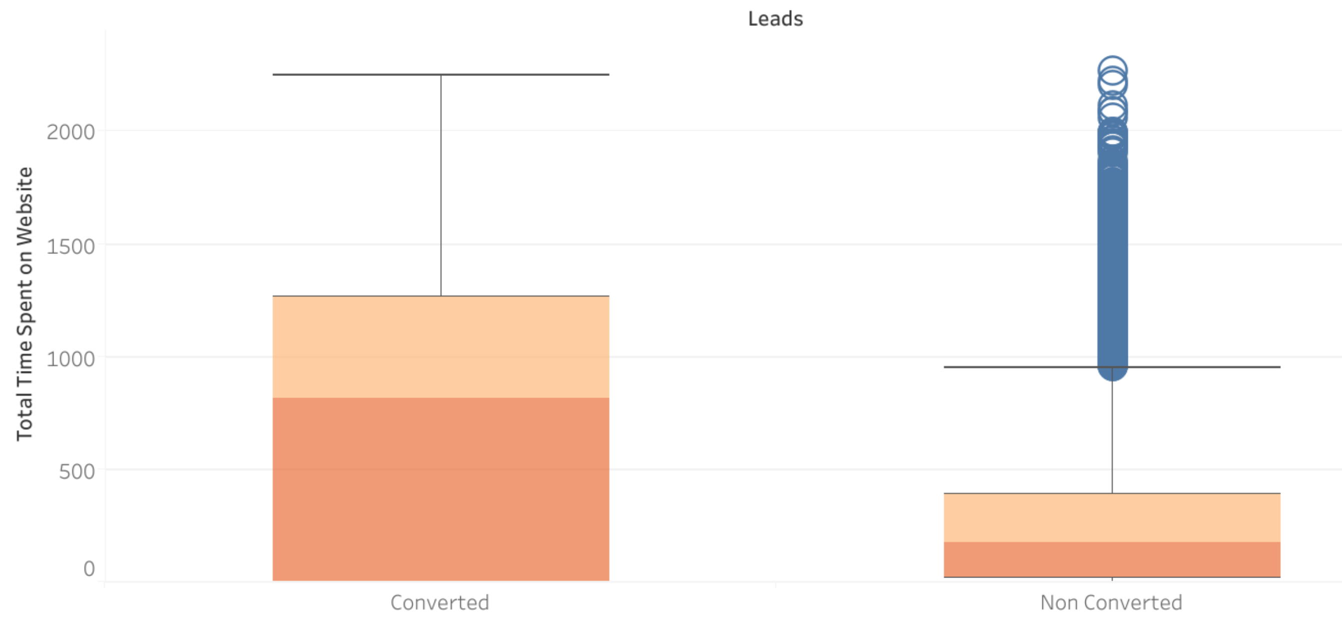


- The group of "Working professionals" appeared to have the greatest conversion rate overall, which is pretty evident, followed by those who were "Unemployed."

OUTLIER ANALYSIS

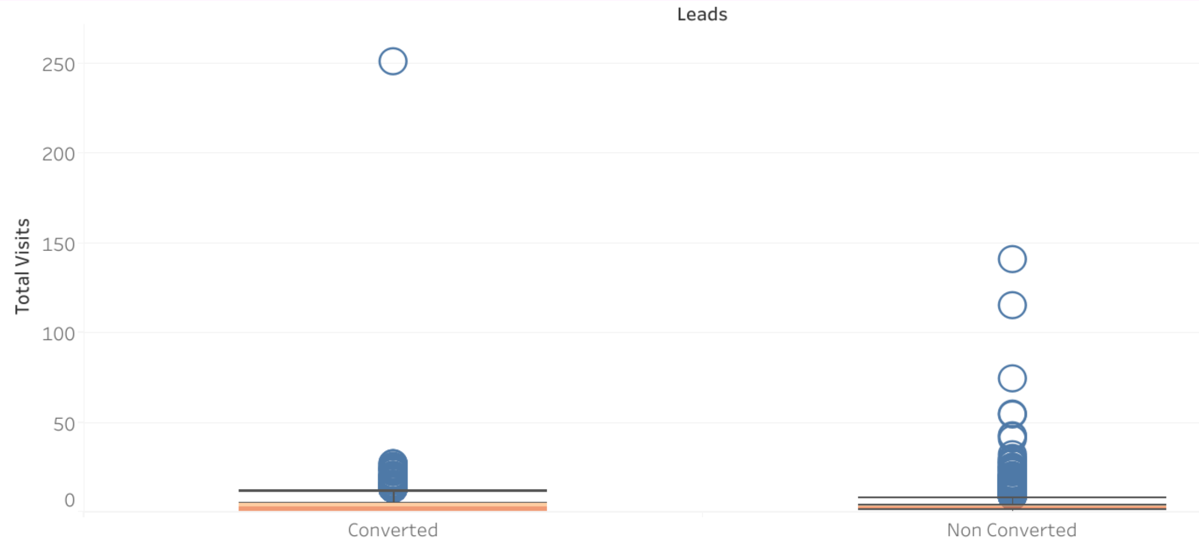
TOTAL TIME SPENT ON WEBSITES VS CONVERTED

The boxplot highlights the fact that clients who have spent more time on the website are more likely to convert, compared to those who have spent lesser time on the site and we don't see any abnormality in the 'Total time spent on website' distribution.

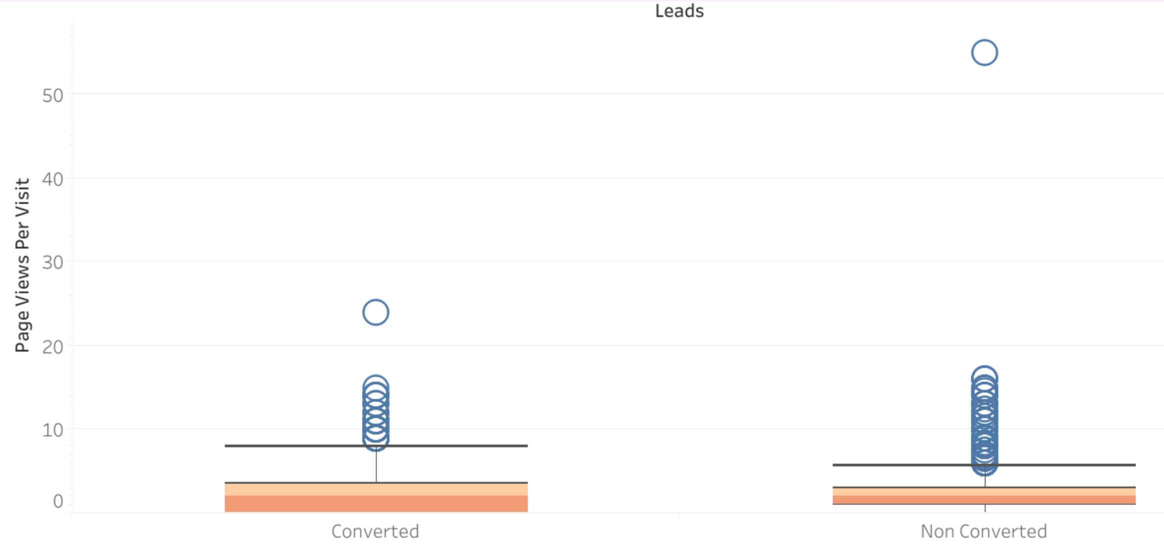


OUTLIER ANALYSIS

TOTAL VISITS VS CONVERTED



PAGE VIEWS PER VISIT VS CONVERTED

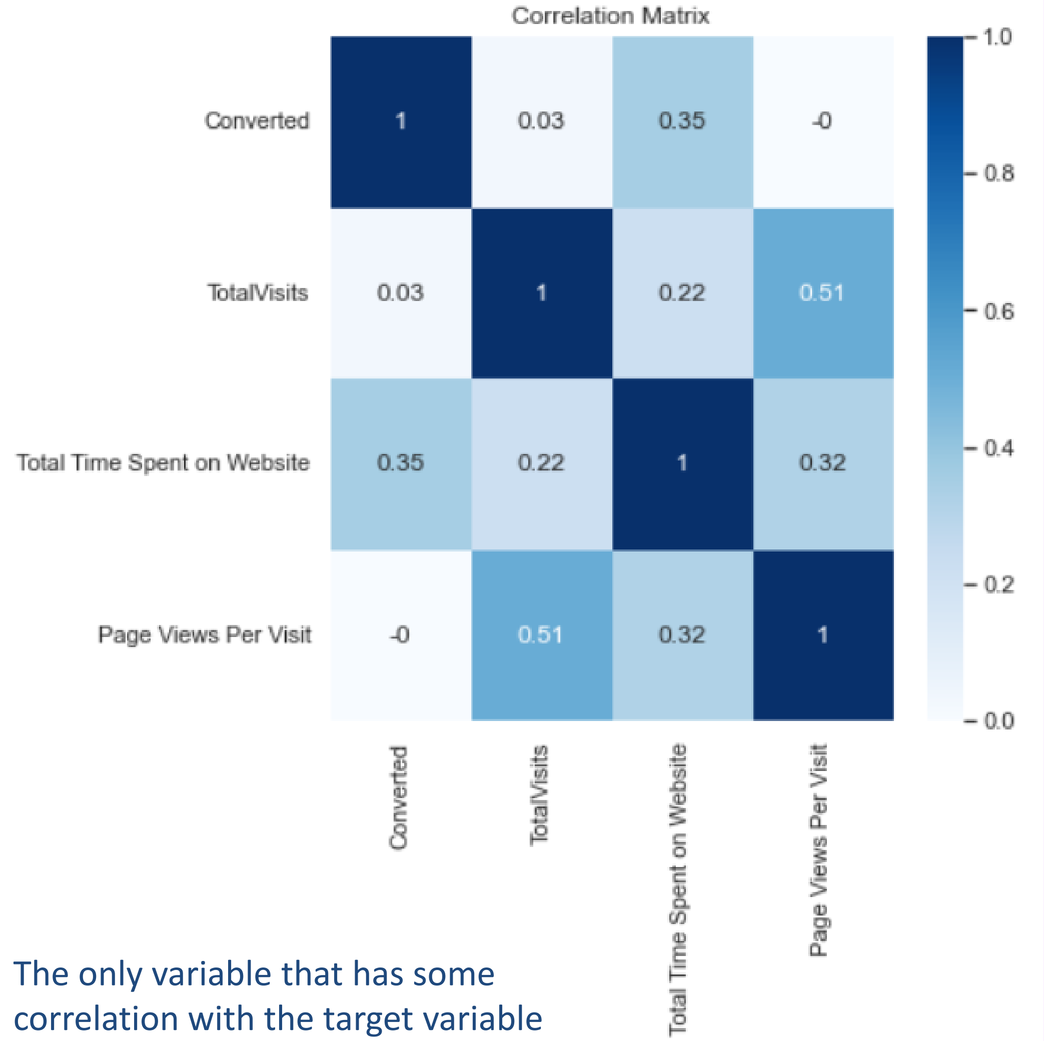


Outliers are present for both the variables "Total visits" and "Page views per visit," as the boxplot shows. Therefore we are not considering these two variable for our model building as though they are not having a significant impact on our target variable.

MODEL BUILDING

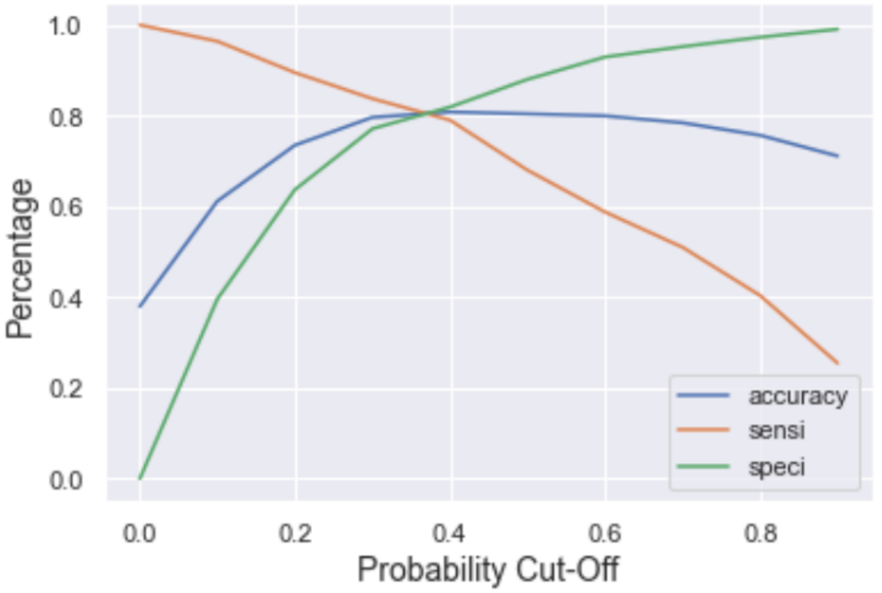
- Splitting the Data into Training and Testing Sets
- Feature Scaling using Standard Scalar
- Model Building using Logistic Regression
- Using RFE to eliminate insignificant variables
- Eliminating variables with high p-values and VIF values
- Prediction on the train set
- Evaluating Model using performance metrics
- Prediction on the test set

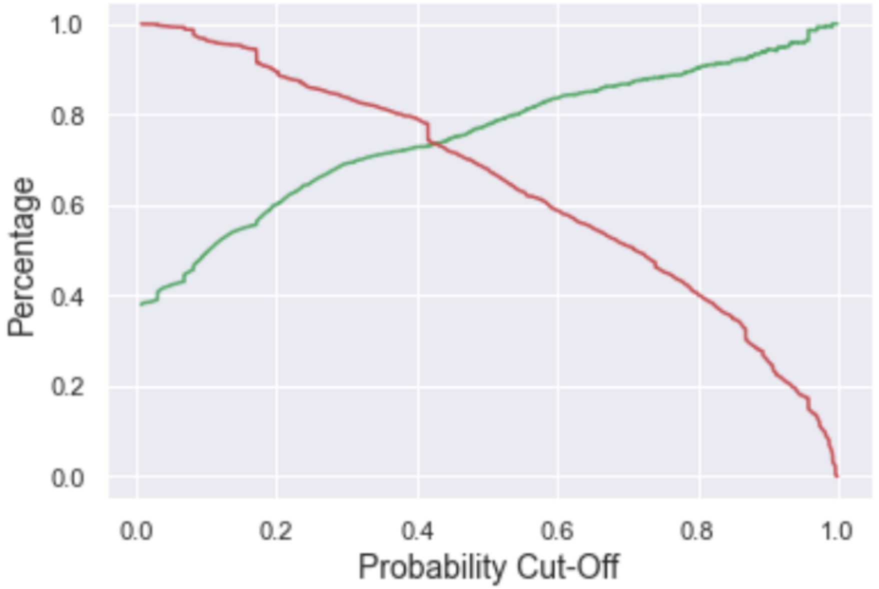
Correlation Matrix



The only variable that has some correlation with the target variable is "Total Time Spent on Website".

MODEL EVALUATION (TRAIN SET)





ACCURACY, SENSITIVITY AND SPECIFICITY

Confusion Matrix

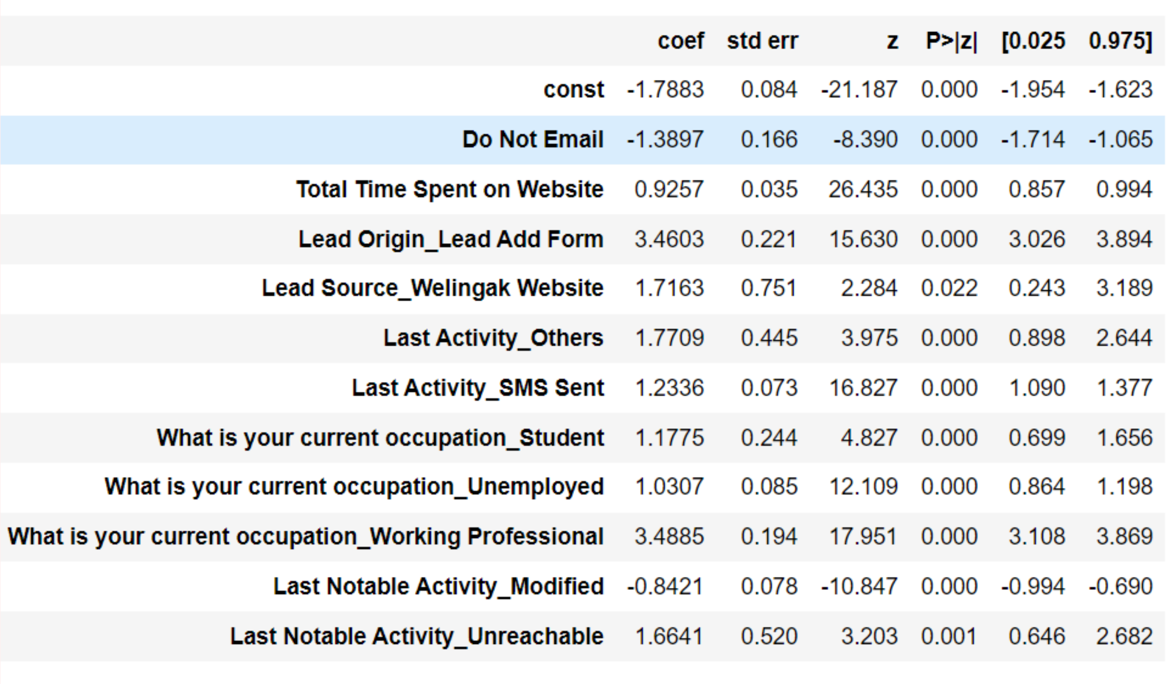
3191	762
477	1942

- ✓ Accuracy - 80.5%
- ✓ Sensitivity - 80.2%
- ✓ Specificity – 80.7%

PRECISION AND RECALL

- ✓ Precision - 77.2%
- ✓ Recall - 77.6%

MODEL EVALUATION (TEST SET)



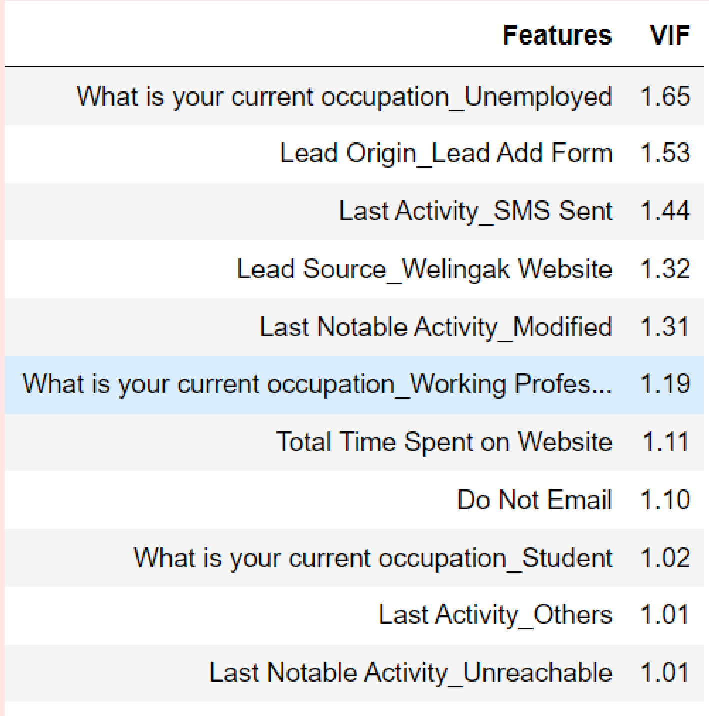
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.7883	0.084	-21.187	0.000	-1.954	-1.623
Do Not Email	-1.3897	0.166	-8.390	0.000	-1.714	-1.065
Total Time Spent on Website	0.9257	0.035	26.435	0.000	0.857	0.994
Lead Origin_Lead Add Form	3.4603	0.221	15.630	0.000	3.026	3.894
Lead Source_Welingak Website	1.7163	0.751	2.284	0.022	0.243	3.189
Last Activity_Others	1.7709	0.445	3.975	0.000	0.898	2.644
Last Activity_SMS Sent	1.2336	0.073	16.827	0.000	1.090	1.377
What is your current occupation_Student	1.1775	0.244	4.827	0.000	0.699	1.656
What is your current occupation_Unemployed	1.0307	0.085	12.109	0.000	0.864	1.198
What is your current occupation_Working Professional	3.4885	0.194	17.951	0.000	3.108	3.869
Last Notable Activity_Modified	-0.8421	0.078	-10.847	0.000	-0.994	-0.690
Last Notable Activity_Unreachable	1.6641	0.520	3.203	0.001	0.646	2.682

ACCURACY, SENSITIVITY AND SPECIFICITY

Confusion Matrix

1373	316
206	836

- ✓ Accuracy - 80.8%
- ✓ Sensitivity - 80.2%
- ✓ Specificity – 81.2%



Features	VIF
What is your current occupation_Unemployed	1.65
Lead Origin_Lead Add Form	1.53
Last Activity_SMS Sent	1.44
Lead Source_Welingak Website	1.32
Last Notable Activity_Modified	1.31
What is your current occupation_Working Profes...	1.19
Total Time Spent on Website	1.11
Do Not Email	1.10
What is your current occupation_Student	1.02
Last Activity_Others	1.01
Last Notable Activity_Unreachable	1.01

This is our final model as all p-values and VIF values are in range now and seems to have no multicollinearity

CONCLUSION

Focus:

- Targeting and contacting website visitors who spend more time than usual there might generate promising leads beneficial for conversions.
- Clients with the Last Activity of "SMS sent" have a better probability of converting.
- Working professionals have a very high conversion rate, hence can be very promising leads.
- The "Lead Add Form" is significant for the sales team since it indicates that the conversion rate for this variable is 93%.
- Due to the high rate of conversion for clients who access the "Welingak Website," it should be given greater attention.
- Marketing strategies need to be reorganized for clients that are unemployed.