Summary

The aim is to build a logistic regression model which can achieve a target lead conversion rate to be around 80%.

Initial Steps and EDA

The solution began with analysis of the percentage of null values in the columns. The columns which had **more than 55% of null values and hence, were instantly dropped.** The columns with lower percentage of null values were kept with the condition of being significant. We would proceed with imputation of values in their case, else we would drop them. The columns such as 'asymmetrique_activity_score', 'asymmetrique_profile_score', etc. were dropped as we could derive any significance based on the unique value analysis. Some others such as city, country, etc. which were considered significant and thus, were preserved with imputation procedure. Further, we performed data visualization with univariate, bivariate and multivariate analysis. The target variable was the conversion variable. This was used intensively while performing bivariate analysis.
During the univariate analysis of the numerical variable page_views_per, we found the significant presence of outliers. This demanded us to remove the outliers by removing the top 1% of the data contained in the column.
Lastly, the columns with very few null values were treated by removing the rows which contained null values. We were able to retain 97% of the data we were assigned. The data processing step came to a closure with the addition of dummy variables.

Model Building

We began by splitting the data into train and test data, wherein 70% of the data fell into the train data and the remaining data were to be treated as test data. The numerical columns were treated with min-max scaling technique. Further, we employed the recursive feature elimination to obtain top 20 significant variables. This followed by application of VIF and p-value methods to manually remove the insignificant feature.

Model Evaluation

The model built was applied to obtain probability values which were supplied with different cut-off probabilities. We began with an arbitrary value of 0.5, then, we obtained an optimal cut-off value of 0.37 using accuracy, sensitivity and specificity plot. Additionally, we also obtained a cut-off value of 0.41 using precision-recall tradeoff curve.

The summary of values of metrics obtained on train data are as follows:

Metrics\Cut-off	**0.5**	**0.37**	**0.41**
Accuracy	0.8488	0.8495	0.8544

Sensitivity(recall)	0.7452	0.8288	0.8136
Specificity	0.6697	0.6329	0.6418
Precision	0.8352	0.7835	0.8021
F-Score	0.7876	0.8055	0.8078

The summary of values of metrics obtained on test data are as follows:

Metrics\Cut-off	**0.37**	**0.41**
Accuracy	0.8443	0.8495
Sensitivity(recall)	0.8362	0.8197
Specificity	0.6211	0.6308
Precision	0.7746	0.7936
F-Score	0.8042	0.8064

Conclusion

Top variables contributing to the conversion are as follows:

1. totalvisits
2. total_time_spent
3. lead_source: olark_chat, reference, welingak_website.
4. lead_quality: not_sure, might_be, worst, low_in_relevance
5. last_activity: sms_sent
6. last_notable_activity: olark_chat_conversation, unreachable
7. do_not_email_yes
8. asymmetric_activity_index_low

The model has achieved an accuracy of about 85% and hence, has sufficiently met the business goal.