SUMMARY

The objective of the case study is to select the most promising leads ie hot leads and to build a logistic regression model & to assign a lead score value to each of the leads.

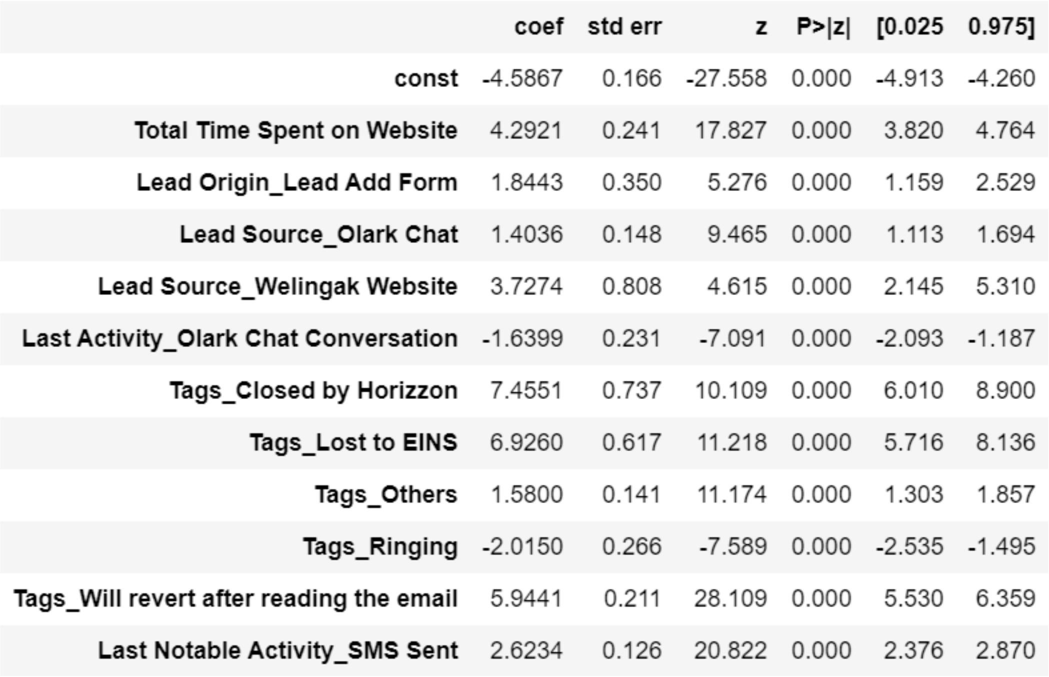
We started with data cleaning which included replacing ‘Select’ with NaN as nothing selected is equivalent to null value, deleting columns having null value > 45 %, handling outliers, imputing the missing values with mode (for categorical variables) and with median (for continuous variables).

In data preparation, we mapped binary variables with 0 & 1 for No & Yes respectively. We have also created dummy variables for multilevel categorical variables.

We splitted the data in Train and Test data in the ratio of 70 : 30.

We have used min-max scaler for standardizing the continuous variables in test data. We started building the model with 15 variables using RFE and done iterations on it till we got the required model.

Below is the final model having 11 variables:



	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.5867	0.166	-27.558	0.000	-4.913	-4.260
Total Time Spent on Website	4.2921	0.241	17.827	0.000	3.820	4.764
Lead Origin_Lead Add Form	1.8443	0.350	5.276	0.000	1.159	2.529
Lead Source_Olark Chat	1.4036	0.148	9.465	0.000	1.113	1.694
Lead Source_Welingak Website	3.7274	0.808	4.615	0.000	2.145	5.310
Last Activity_Olark Chat Conversation	-1.6399	0.231	-7.091	0.000	-2.093	-1.187
Tags_Closed by Horizzon	7.4551	0.737	10.109	0.000	6.010	8.900
Tags_Lost to EINS	6.9260	0.617	11.218	0.000	5.716	8.136
Tags_Others	1.5800	0.141	11.174	0.000	1.303	1.857
Tags_Ringing	-2.0150	0.266	-7.589	0.000	-2.535	-1.495
Tags_Will revert after reading the email	5.9441	0.211	28.109	0.000	5.530	6.359
Last Notable Activity_SMS Sent	2.6234	0.126	20.822	0.000	2.376	2.870

We then predicted the probability of dependent variable i.e ‘Converted’ for the train data set.

We also predicted the conversion probability on the cut-off value of at 0.5 ie leads having prob. > 0.5 are predicted as converted and others as 0 on train data set.

From sensitivity, specificity and precision curve we got cut-off at 0.25 and then we predicted the target variable on the basis of cut off value i.e leads having prob. > 0.25 are predicted as converted and others as 0. We have also calculated the accuracy, sensitivity, specificity, precision etc using confusion matrix.

We have then used precision-recall curve to identify the cut-off prob and got the cut off as 0.35. We found that the model was giving same results at cut-off 0.25, hence we predicted the conversion of test data at the same cut-off.

We, then evaluated the model on the test data set and calculated the evaluation metrics.

Following is the result of metric on the test and train data:

	Train	Test
Accuracy	91.10 %	91.29 %
Sensitivity	90.61 %	91.52 %
Specificity	91.40 %	91.16 %
Precision	86.52 %	86.66 %
F1 Score	88.52 %	89.02 %

F1 Score of the built model for both train and test data is good.

For calculating the lead score we have merged both the test and train data frame.

Lead score = 100 * Conversion probability.

Higher the lead score, higher is the probability of a lead getting converted and vice versa.

In the final model, we got 11 features.

Top three features which lead to higher conversion of the leads are:

- Tags_Closed by Horizzon
- Tags_Lost to EINS
- Tags_Will revert after reading the e-mail