ML – Case Study

This case study aims to identify the 'Hot Leads' - the most potential leads , i.e. the leads that are most likely to convert into paying customers. And to build logistic regression a model to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

Steps followed:

1. EDA on the dataset. Read the data, find missing values and do the Outlier Analysis.

   Problems faced: There were so many missing values, few columns had invalid value 'Select' which is equal to null value. Few binary categorical columns did not have enough data for one of the categories (eg: more than 95% was NO less than 2% was Yes).

   Way Handled:
   a) 'Select' is converted to Nan value and then all the Nan are imputed with either mode or random value from that column based on the distribution of that variable.
   b) Dropped unnecessary columns, columns with >30% data unavailable, columns with <2% missing data – rows with missing data is removed.
   c) There are certain columns which we thought are important from business sense but had more then 50 % null values

like Lead Quality /Lead Profile but these are not dropped instead these are imputed with the either mode or with random values.

d) did univariate analysis to check data distribution.
e) Handled outliers by deleting data above 0.95 and below 0.05 quantile.

2. Build the model using Logistic regression.

   Before building model, for binary variables converted ‘Yes’ to ‘1’ and ‘No’ to ‘0’. Created dummy variables. Split data to train and test sets.

   Problems faced: Number of attributes were very high due to the more no of categories in categorical variables and few variables were correlated.

   Way Handled :

   1) Wherever possible combined the less then 10% categories into 1 single category as others
   2) Using RFE feature selection reduced the variables to 13.
   3) Then based on VIF values dropped few variables, built final model where all VIFs are less. Checked conversion matrix, calculated accuracy, specificity, sensitivity etc.
   4) Based on the analysis of ROC curve, cutoff analysis decided to use 0.3 as cutoff value.

3. Test the model on the test data and check the accuracy score. Observed that the Accuracy score on the test (0.89) and train data (0.88) is almost SAME

Finally applied the model on test data.

Final observations:

Following are the variables that have high impact on Lead score.

Main factors.	**Coefficient**
Tags_Will revert after reading the email	4.2012
Lead Origin_LeadOriginOther	3.8368
Lead Quality_High in Relevance	2.4856
Last Notable Activity_SMS Sent	1.8342
Lead Source_Olark Chat	1.4128
Total Time Spent on Website	1.0710
Lead Quality_Low in Relevance	0.9677
Last Activity_LastActivityOther	-0.8411
Last Activity_Page Visited on Website	-0.6047
Do Not Email	-1.0166
Last Activity_Olark Chat Conversation	-1.6466
Lead Quality_Worst	-2.0549
Tags_Ringing	-3.5245

Looking at the coefficients of all the variables we could say the top three variables in our model that contribute more towards the probability of a lead getting converted are Tags, Lead Origin, Lead Quality.