**Lead Scoring Case Study Report**

**220940325067 Shruti Kishor Patil**

**220940325006 Ajay Prakash Fatpure**

**220940325057 Ruchira Rajesh Bhargave**

**220940325032 Gaurav Rajan Bopardikar**

**220940325068 Shubham Shrivastava**

**Summary:**

The first step that we took while performing the case study was to inspect the data. Upon inspection we found that the data had multiple null values. There were still as few columns that needed to be imputed as their null value percentage was not high enough for these to be dropped directly from the dataset. We then imputed all these column values with the highest occurring value in each column. We also dropped ‘Magazine’, ‘Receive More Updates About Our Courses’,’ Update me on Supply Chain Content’, ‘Get updates on DM Content’, ‘I agree to pay the amount through cheque' columns as they had only single values and didn’t serve our analysis. We then handled all the variables which had either Yes or No and converted them to numerical data . We also dropped Search Newspaper ArticleX , Education Forums, Newspaper, Digital Advertisement Through Recommendation. After dropping all these values we created dummy variables for all the categorical variables and dropped the original columns of each. We then divided the dataset into test and train datasets respectively. We then performed logistic regression on this newly picked columns or predictor variables. Here we build the Logistic Regression Model with 90.13% accuracy.

Here are the steps we followed:

**1. EDA:**

• We dropped columns which are not required for analysis like Prospect\_Id.

• Quick check was done on % of null values and we dropped columns with more than 45% missing values.

• We replaced the ‘select’ values in some columns with ‘Nan'.

• We checked the Imbalance data…..wherever we found that , we just dropped those columns as they can be a factor of biased result.

• We also filled some Null values in those columns which are relevant for analysis.

• To get the only desired columns, we had gone through columns one by one, checked its importance and made the decision.

**2. Train-Test split & Scaling :**

# Splitting the data into train and test

from sklearn.model\_selection import train\_test\_split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0

• The split was done at 80% and 20% for train and test data respectively.

• We have done scaling on the variables ['TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website']

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

X\_train[['TotalVisits','Total Time Spent on Website','Page Views Per Visit']] = scaler.fit\_transform(X\_train[['TotalVisits','Total Time Spent on Website','Page Views Per Visit']])

X\_test[['TotalVisits','Total Time Spent on Website','Page Views Per Visit']] = scaler.fit\_transform(X\_test[['TotalVisits','Total Time Spent on Website','Page Views Per Visit']])

**3. Model Building**

#Logistic Model

from sklearn.linear\_model import LogisticRegression

classifier = LogisticRegression(random\_state=1)

model = classifier.fit(X\_train, y\_train)

• We have attained top relevant columns out of which all are independent except 1 which is dependent i.e (‘converted’ column).

• Dummy variables of the categorical features were created by using get\_dummies() method.

• As our target variable is categorical, for providing training or fitting the model to the training set, we will import the Logistic Regression class of the sklearn library.

• After importing the class, we have created a classifier object and used it to fit the model to the logistic regression.

• We also observe the summary.

* Predicting the Test Result Our model is well trained on the training set, so we have predicted the result by using test set data.

**5. Model Evaluation**

• We have created the confusion matrix here to check the accuracy of the classification.

• True Positive = 1099

• True Negative = 537

• False Negative = 91

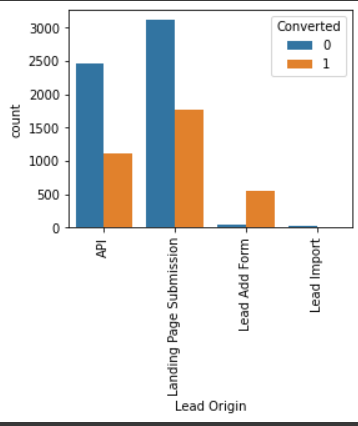
• False Positive = 88

Accuracy = 90.13%

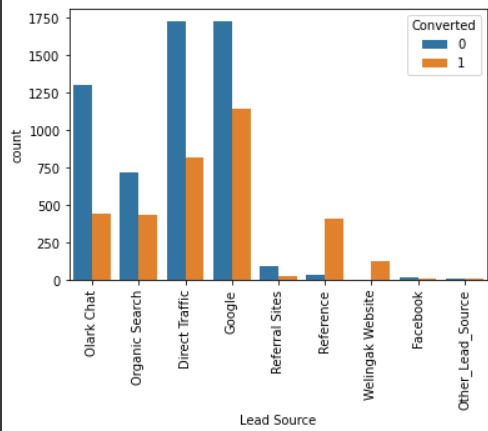
**CONCLUSION**

1. Here we build the Logistic Regression Model with 90.13% accuracy.
2. The top three variables in my model, that contribute towards lead conversion are:

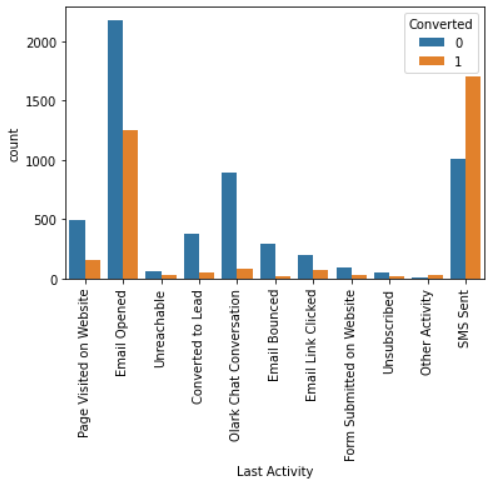
1. Lead Origin



2. Lead Source



3. Last\_Activity



**2. To focus on wider set of lead with –**

Tags: will revert after reading the email

Total Time Spent on Website

Last\_ Activity\_SMS Sent

Tags\_Lost to EINS

Last Activity\_SMS Sent

Current occupation\_Working Professional

