Lead Scoring Case Study

By:

Amarpreet Singh

and

Mathankumar Selvaraj

Date: 07/09/2019

Problem Statement

- To help X education to select the most promising leads known as 'hot leads' who are most likely to convert into paidcustomers.
- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads where the leads with higher lead score have a higher conversion chance and the leads with lower lead score have a lower conversion chance.
- Identify the driver variables and understand their significance which are strong indicators of lead conversion.
- Identify the outliers, if any, in the dataset and justify the same.
- Consider both technical and business aspects while building the model.
- Summarise the conversion predictions by using evaluation metrics like accuracy, sensitivity, specificity and precision.

Data Exploration

- ‘**Leads.csv**’contains all the information about the leads generated through various sources and their activities.
  - This file contains 9240 rows and 37 columns.
  - Out of 37 columns, 7 are numeric columns and 30 are non-numeric or categorical columns.
  - Current conversion rate of the leads is 39%.
- **‘Leads Data Dictionary.csv’**is data dictionary which describes the meaning of the variables present in the “Leads” dataset.

Data Cleaning andPreparation

Leads.csv :

- Following columns contain more than 30% null values initially:
    1. What is your current occupation
    2. Tags
    3. Lead Quality
    4. Asymmetrique Activity Index
    5. Asymmetrique Profile Index
    6. Asymmetrique Activity Score
    7. Asymmetrique Profile Score
- Following columns have default value of ‘select’ as a dominating value which is same as null value. So, we have converted ‘select’ to ‘Others’.
    1. Specialization
    2. How did you hear about X Education
    3. Lead Profile
- All the missing values of categorical columns have been imputed with ‘Mode’.

Data Cleaning andPreparation

- All the missing values of quantitative columns have been imputed with median as the difference between mean and median is insignificant.

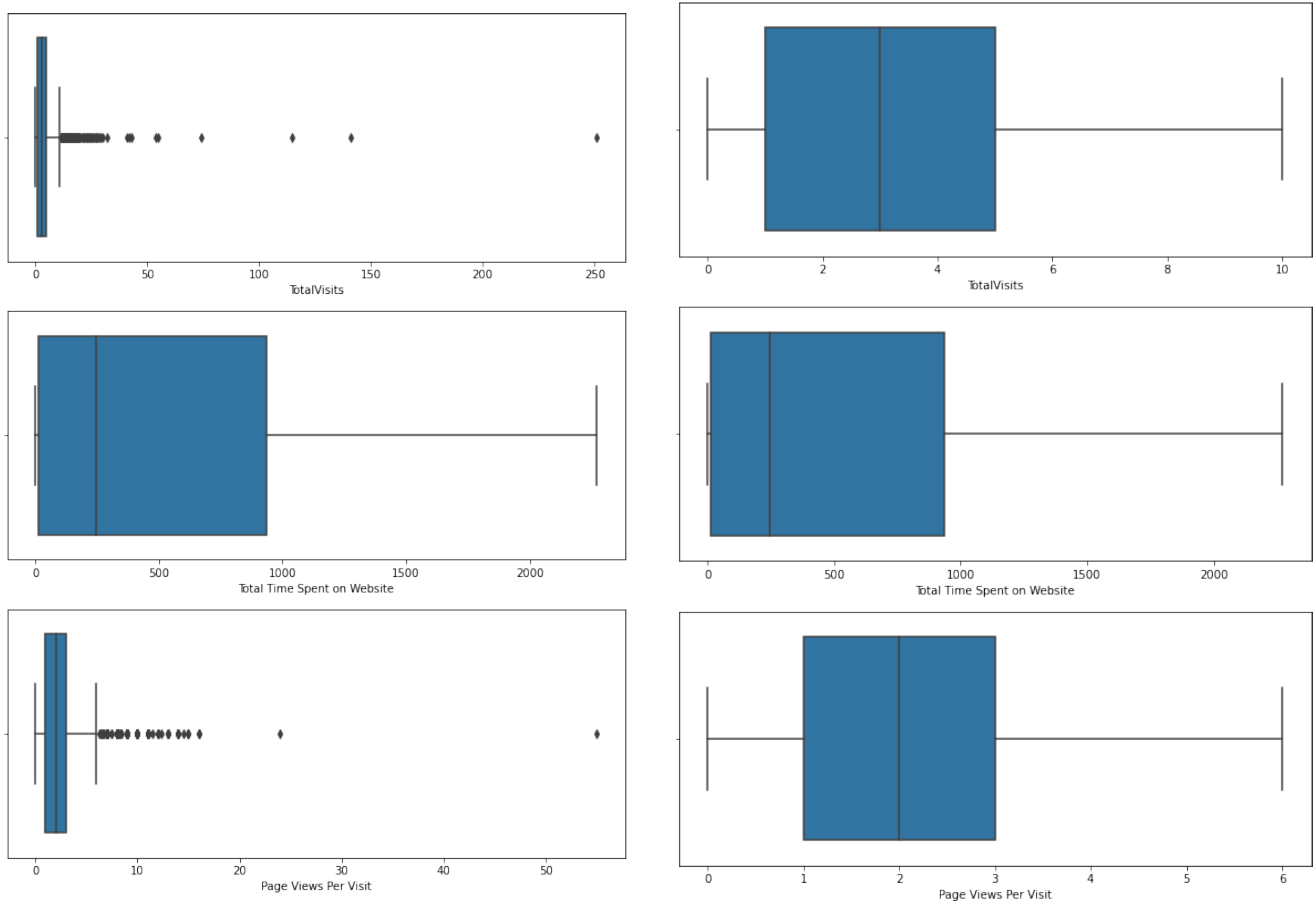
- Following columns have been dropped which contain single value as their contribution is insignificant:
  1. Magazine
  2. Receive More Updates About Our Courses
  3. Update me on Supply Chain Content
  4. Get updates on DM Content
  5. I agree to pay the amount through cheque

- Following columns have been imputed with mode since the percentage of missing value is low.
  1. Lead Source
  2. Lead activity

- Following columns have been dropped as they have either been highly skewed or insignificant for the model building
  1. What matters most to you in choosing a course
  2. Country
  3. City
  4. Prospect ID
  5. Lead Number

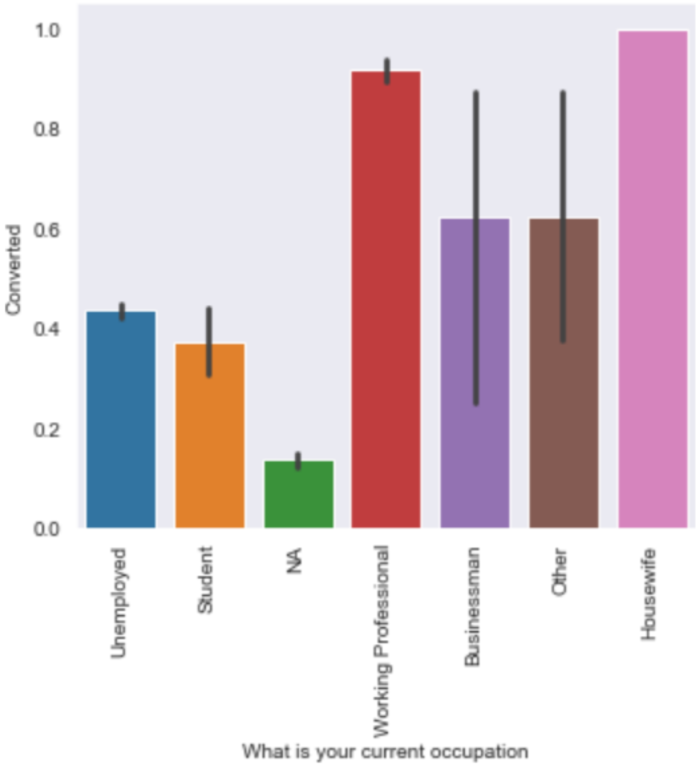
UnivariateAnalysis – Outliers

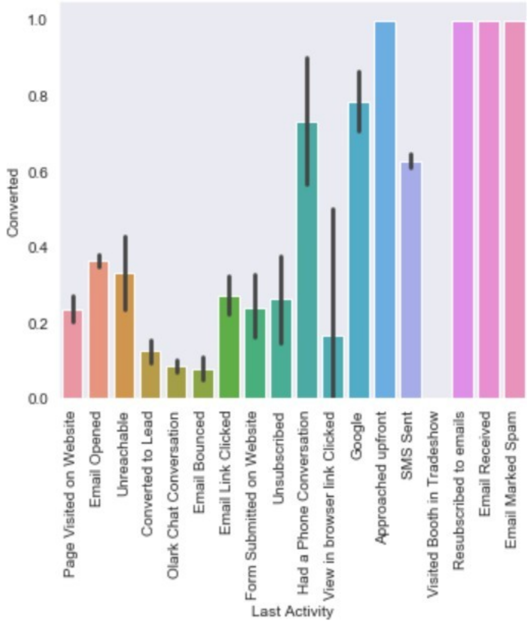
- Univariate analysis revealed data distribution and outliers in ‘Leads’ data. Key columns where outliers were identified are:-
  - a. TotalVisits
  - b. Page Views Per Visit
  - c. Total Time Spent on Website
- Using 95 percentile to mitigate the effects of outliers.
- Decision has been taken to not remove any outlier.
- We will review the final model to ensure this does not impact the score.

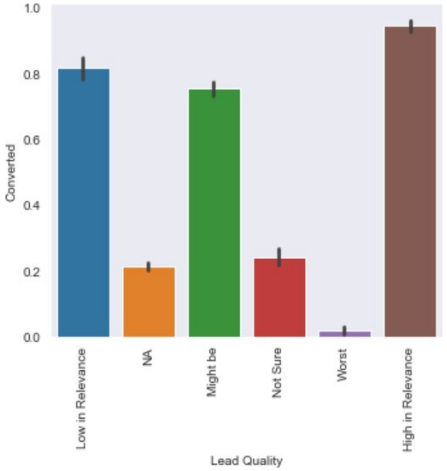


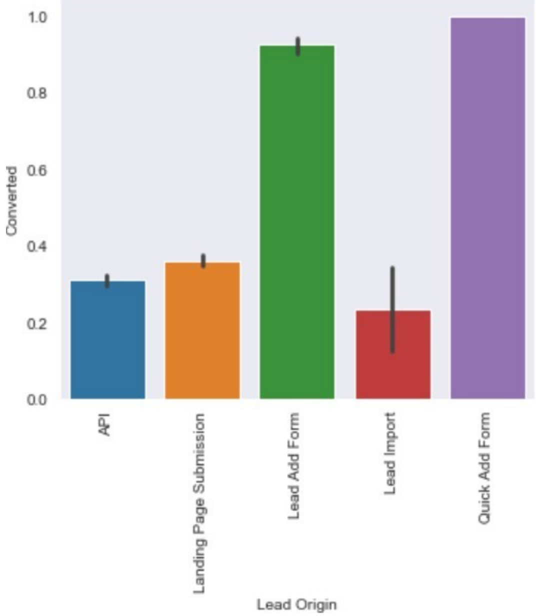
BivariateAnalysis:Categorical Variables

- ‘Converted’ column has been chosen as target variable. So, bivariate analysis of important variables has been performed with respect to the target variable.
- Lateral students and the visitors showing interest on next batch have higher chances of getting converted.
- Lead quality tagged with “High in Relevance” has high conversion rate history.
- Lead originated through “Lead Add Form” and “Quick Add Form” has high possibility of getting converted.
- Lead belongs to Welingak Website, WeLearn, Live Chat and NC_EDM converts more than any other sources.









Data Preparation for Modelling

Create Dummy Variables:

- Independent variables as dummy variables allows easy interpretation and calculation of the odds ratios, which increases the stability and significance of the coefficients.
- Dummy variables have been created for following columns:
  1. Lead Origin
  2. Lead Source
  3. Last Activity
  4. Specialization
  5. What is your current occupation

Label Encoding:

- Label encoding is simply converting each value in a column to a number.
- We will use label encoding for variables with higher level. This is to avoid drastic increase in data-frame size.
- All the relevant categorial variables have been encoded using ‘LabelEncoder’.

Data Preparation for Modelling

Binary Variables Encoding:

- Variables which have binary (Yes/No) values have been encoding with 1 and 0.
- 1 denotes Yes whereas 0 denotes No.

Train – Test Split:

- The modified 'Leads' dataset has been split into Train and test dataset in the ratio 70:30.
- Train dataset has been used to train the model whereas Test dataset has been used to evaluate the model

Feature Scaling:

- It is important to have all variables on the same scale in order to avoid the dominance of variables with high magnitude in the model.
- "StandardScaler" function has been used to scale the data for modeling which brings all the data points into a standard normal distribution with mean at '0' and standard deviation at '1'.

Model Building: Using logistic Regression

- Generalised Linear Model (GLM) from StatsModels library has been used to build the Logistic Regression Model.
- Initially, the model was built using 93 features present in X_train dataset.
- Most of the features were found to be insignificant. So, we needed to perform feature selection technique.

Feature Selection using Recursive Feature Elimination (RFE):

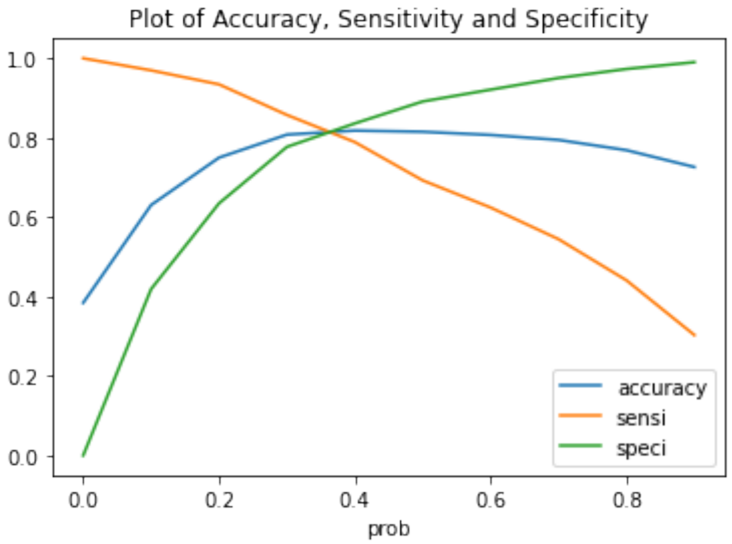
- **RFE** is an optimisation technique for finding the best performing subset of features. It is based on the idea of repeatedly constructing a model and choosing either the best (based on coefficients), setting the feature aside and then repeating the process with the rest of the features. This process is applied until all the features in the dataset are exhausted. Features are then ranked according to when they were eliminated.
- We ran RFE to identify top 20 features for further model building process.
- Insignificant features were dropped one by one after checking the P-value and Variance Inflation Factor (VIF).
- Accepted P-value should be kept below 0.05 and VIF should be less than 5.

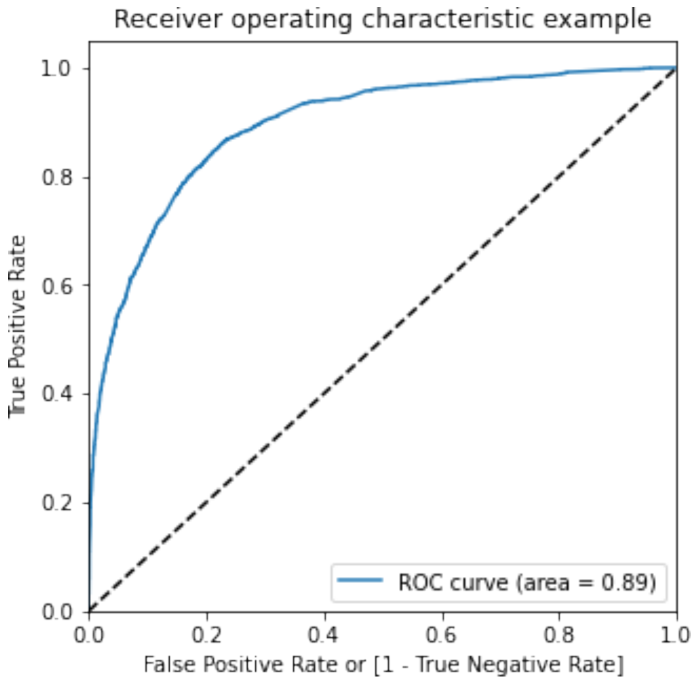
Final Model and Interpretation

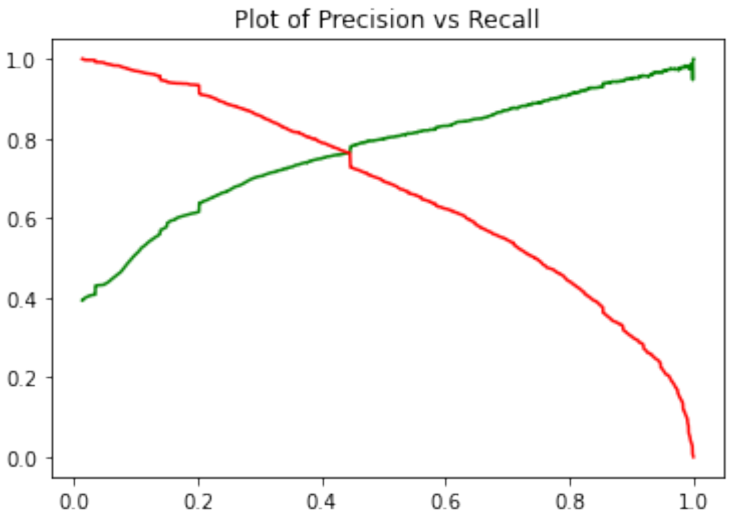
- Final model contains 15 most important features which satisfy all the selection criteria.
- Lead score having conversion probability greater than 0.43 are being predicted as “Converted”.
- Using this probability threshold value (0.43), the leads from the test dataset have been predicted whether they would get converted or not.
- Confusion matrix with cut-off 0.43 has been created to calculate evaluation metrics.
- Confusion matrix:**[1203, 209],**
  [172, 726]
- Evaluation metrics:
  - **Accuracy: 83.50%**
  - **Sensitivity: 80.84%**
  - **Specificity: 85.19%**
  - **Precision: 77.64%**

EvaluationMetrics

- Receiver Operating Characteristics (ROC) Curve:
  - By determining the Area Under the Curve (AUC) of the ROC curve, the goodness of the model is determined.
  - Since the ROC curve is close to the upper left part of the graph, it means this model is a very good model.
  - The value of AUC for our model is 0.89.
- Plot accuracy sensitivity and specificity:
  - Tradeoff between sensitivity and accuracy can be observed (cutoff = 0.38).
- Precision and Recall plot:
  - Ideal cutoff of 0.43 is observed from recall and precision plot.
- We will use both the cutoff and evaluate results for further predictions.


Plot of Accuracy, Sensitivity and Specificity


Receiver operating characteristic example


Plot of Precision vs Recall

Conclusion and Recommendations:

- Followings are top three features that contribute to decision which mean the conversion probability of a lead increases with increase in values of these features:
  - Lead Origin
  - What is your current occupation
  - Last Activity
- Top three categories that contribute to decision
  - Lead Origin ==> Lead Add Form
  - What is your current occupation ==> Working Professional
  - Last Activity ==> SMS Sent

Conclusion and Recommendations:contd..

- This model will help to identify the hot leads which would enhance **speed-to-lead** and the **response rate**.
- Approaching only to hot lead would result in:
- Shorter **sales cycle** through intuitive prioritization.
  - Better **opportunity-to-deal ratio**
  - Control over volatile **buying cycle**
  - Increase **marketing effectiveness**
  - Better **sales forecasting**
  - Minimize opportunities loss
  - Increase in revenue