LEAD SCORE CASE STUDY

Group Members – Ashwini Reddy J & Shesh Mani Tripathi

Problem Statement

- To help X education to select the most promising leads known as ‘hot leads’ who are most likely to convert into paid customers.
- 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.
- Summarize 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.
    i. This file contains 9240 rows and 37 columns.
    ii. Out of 37 columns, 7 are numeric columns and 30 are non-numeric or categorical columns.
    iii. 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

Solution Methodology

- Data cleaning and data manipulation

1. Check and handle duplicate date .
2. Check and handle NA values & missing values.
3. Drop columns, if it contains large amount of missing values & not useful for the analysis.
4. Imputation of the values, if necessary.
5. Check & handle outliers in data .

. **EDA**

1. Univariate data analysis : Value count, distribution of variable etc.
2. Bivariate data analysis : correlation coefficients & pattern b/w the variables etc.

Data Cleaning and Preparation

- Leads.csv :
- Following columns contain more than 30% null values initially:

1. What is your current occupation
2. What matters most to you in choosing a course
3. Tags
4. Lead Quality
5. Lead Profile
6. Asymmetrique Activity Index .
7. Asymmetrique Profile Index .
8. Asymmetrique Activity Score .
9. 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 'NA'.

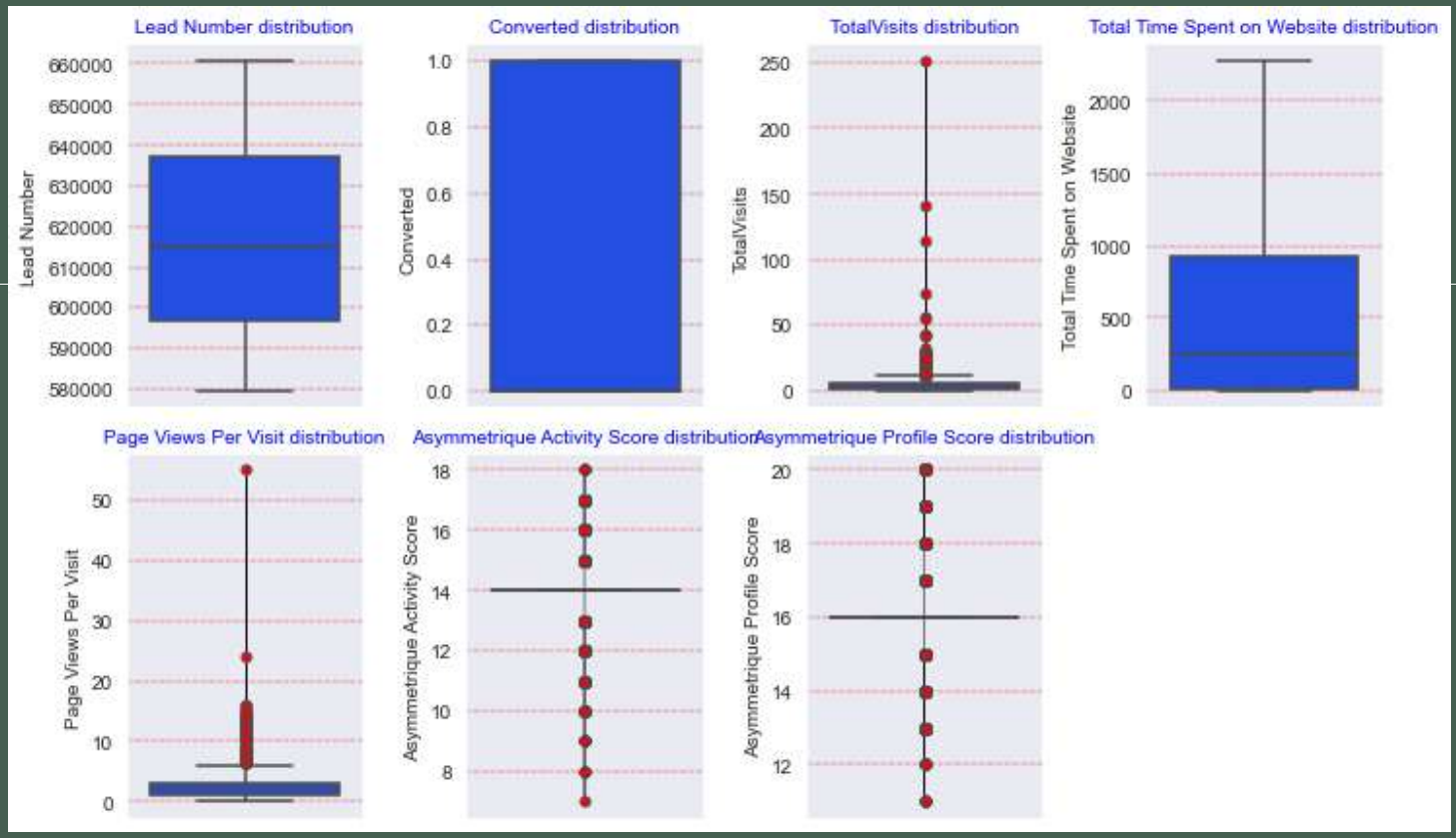
1. Specialization .
2. How did you hear about X Education .
3. Lead Profile .
4. City

- All the missing values of categorical columns have been imputed with 'NA'

Data Cleaning and Preparation

- 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 dropped since percentage of missing value is more than 70%:
  1. How did you hear about X Education
  2. Lead Profile
- Following columns have been imputed with mode since the percentage of missing value is low.
  1. Lead Source
  2. Lead activity

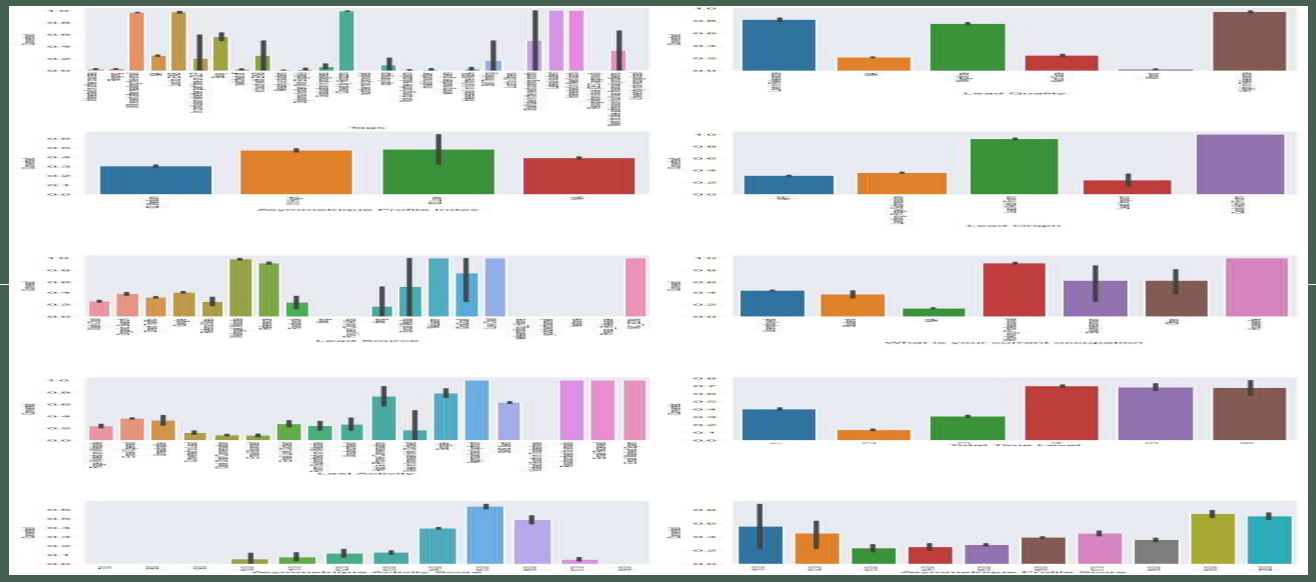
Boxplot for quantitative varibles



Bivariate Analysis: 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.

Bivariate Analysis: Categorical variables



Data Conversion

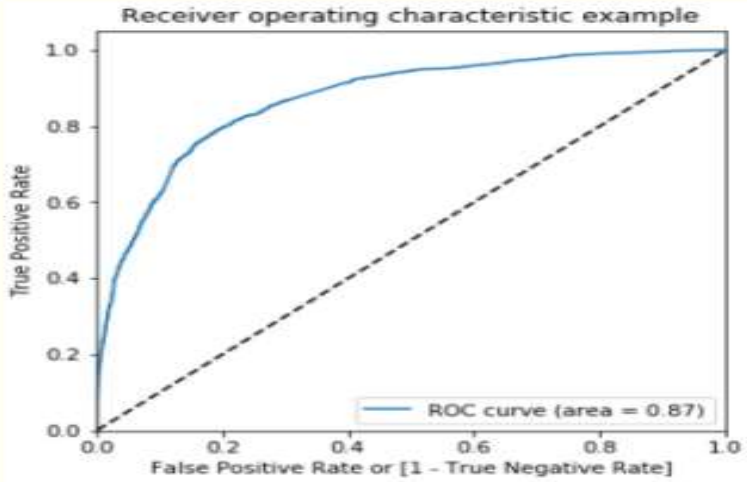
- Numerical Variables are normalised.
- Dummy Variables are created for object type variables.
- Total Rows for analysis :8792
- Total Columns for analysis : 43

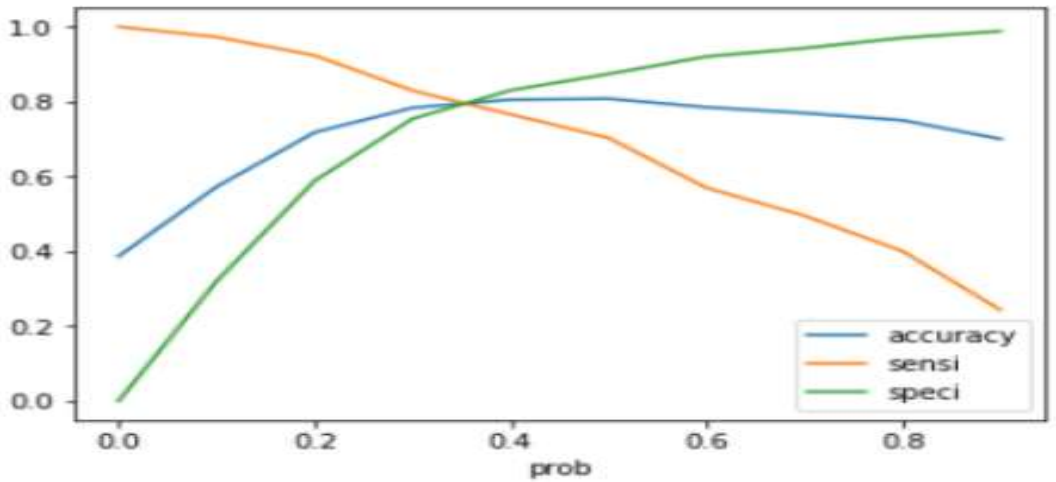
Model Building

- Splitting the data into training and testing sets.
- The first basic step for regression is performing a train- test split, we have chosen 70:30 ratio.
- Use RFE for feature Selection.
- Running RFE with 15 variables as output.
- Building model by removing the variable whose p-value is greater then 0.05 and VIF value is greater than 5.
- Predictions on test data set.
- Overall accuracy 81%

ROC Curve

- Finding optimal cut of point.
- Optimal cut off probability is that
- Probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.35.





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