Lead Scoring Case Study

By –

Susandeep Ganta

Madhur Sharma

Problem Statement

- Education company named X Education sells online courses to industry professionals.
- Company markets its courses on several channels. People fill up a form providing their details, they are classified to be a lead. The typical lead conversion rate is around 30%.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.

Case Study Steps

- Data cleaning and Data imputation
  1. Handling NULL Values.
  2. Drop columns, if it contains large amount of missing values and not useful for the analysis.
  3. Imputation of the values, if necessary.
  4. Checking for Outliers and handling them.
- Feature standardization and data preparation
- Model Building: logistic regression
- Model Evaluation
- Conclusions and Insights

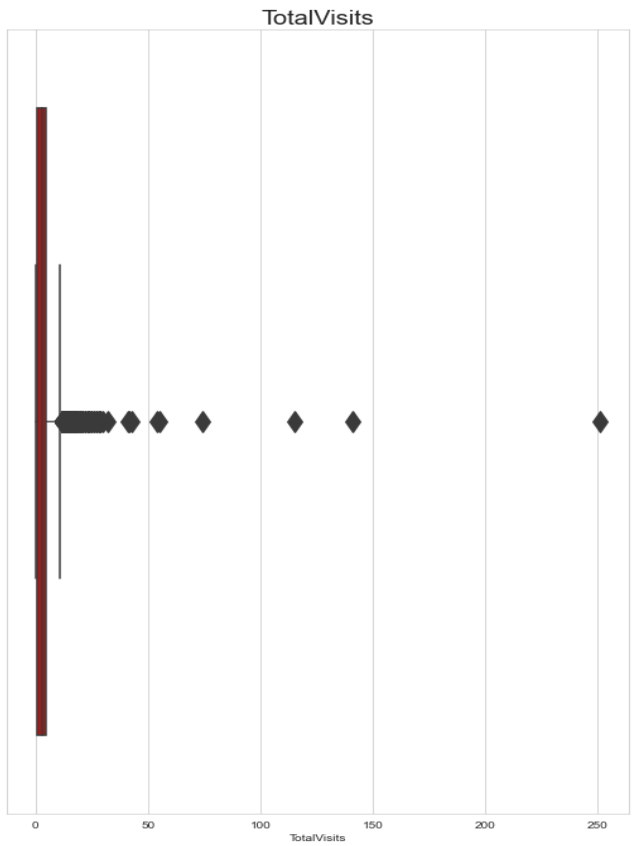
Data cleaning and Data imputation

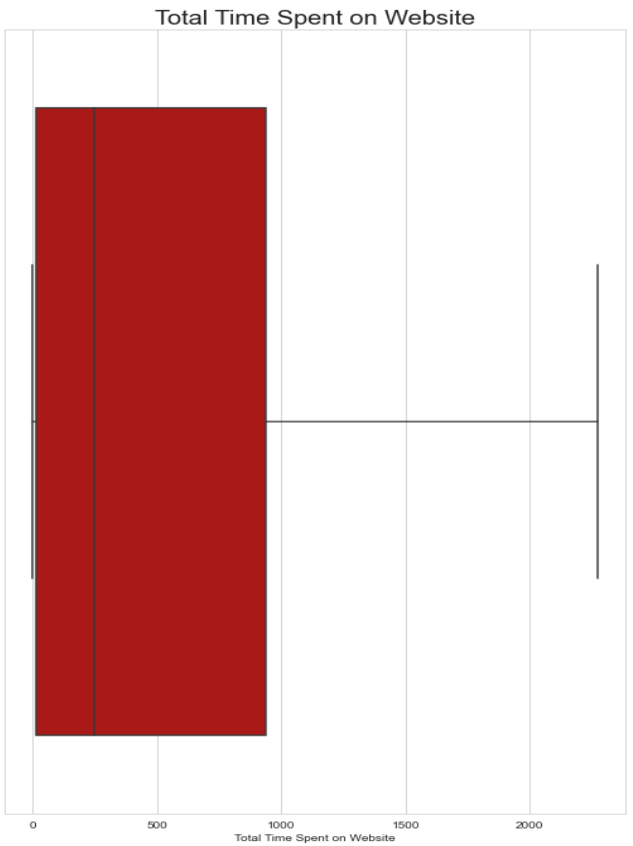
- We observe that out of 37 columns 17 columns have null values. We have decided to remove all the columns having null value greater than 30%. Therefore we'll be dropping 6 columns in total. The other columns which has null value percentage lesser than 30% we'll impute them either with Mean or mode values.
- We see there are couple of columns with null values. We have to find the data type of those columns and then find the type of column it is. If the column is numerical then we can impute the missing data with mean value, if the column has categorical value then we have to replace it mode value.
- We observe that there are couple of columns which has 'select' as the data in it. Hence, we are required to replace those values with the Nan values.

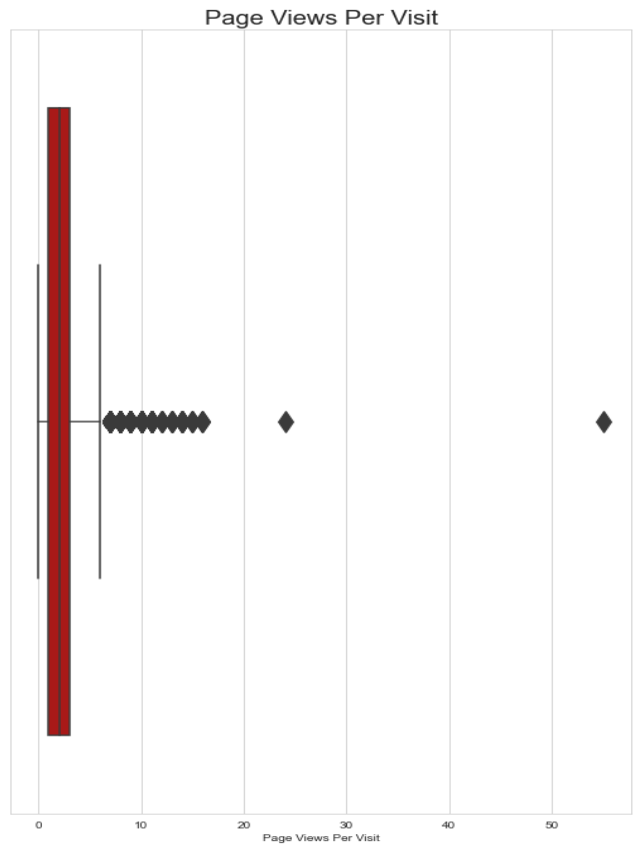
Highlights –

- ✓ After dropping the columns with null value more than 30% we have 22 columns.
- ✓ In those 22 columns we see that there are 5 columns where null value exists.
- ✓ To treat these we must find whether these are categorical columns or numerical columns.
- ✓ Dividing them to categorical and numerical set.
    - ✓ Categorical column:
        - ✓ 'What matters most to you in choosing a course'
        - ✓ What is your current occupation
        - ✓ Last Activity
        - ✓ Lead Source
    - ✓ Numerical column:
        - ✓ Page Views Per Visit
- ✓ We will treat all the missing values of the categorical columns using mode and missing values of the numerical columns with mean.

Outlier Analysis -







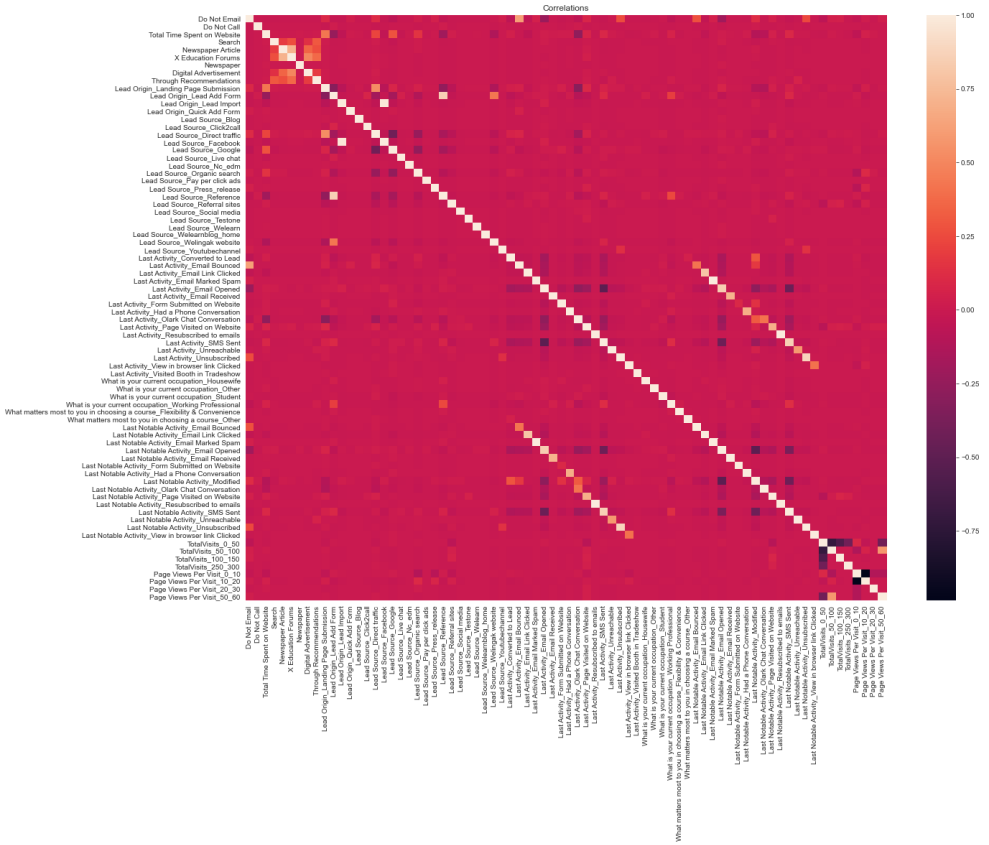
Model Building

- Data Preparation - Train- Test split the data
- Feature scaling
- Checking the correlation
- Building model on train dataset
- Using RFE to select best variables
- Calculating the VIF
- Predicting the test model

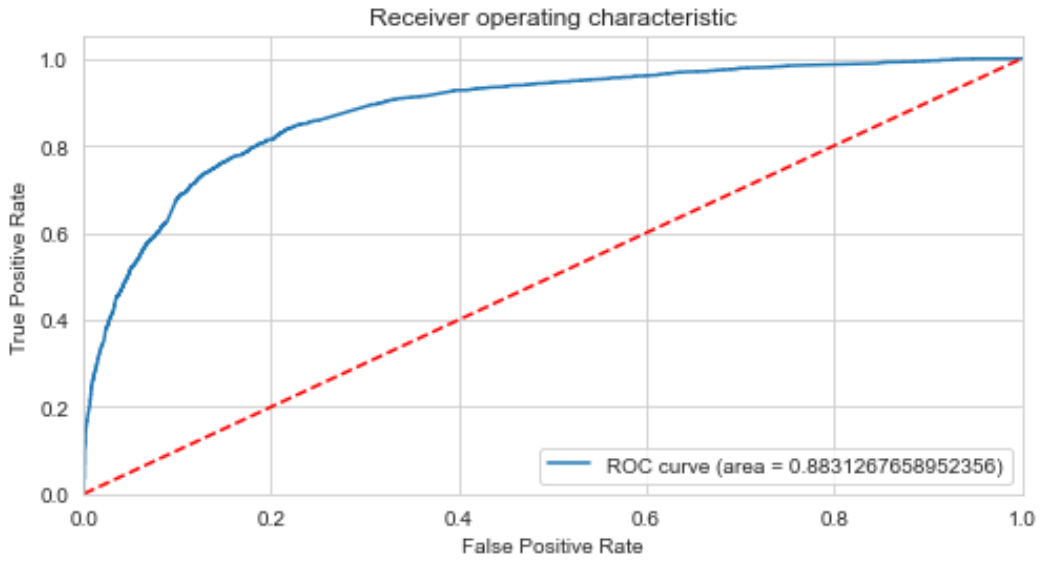
Highlights –

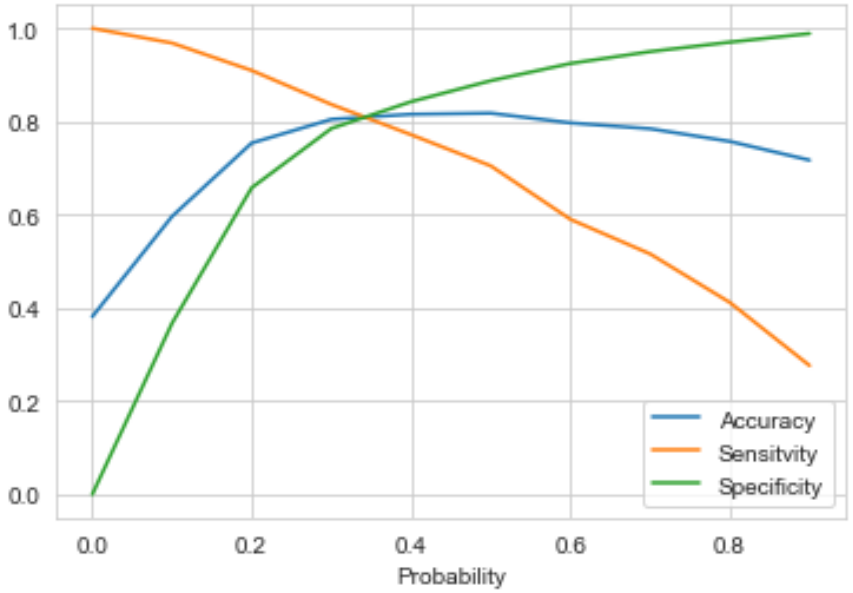
OverAll Accuracy: 82%

Correlation analysis



ROC Curve

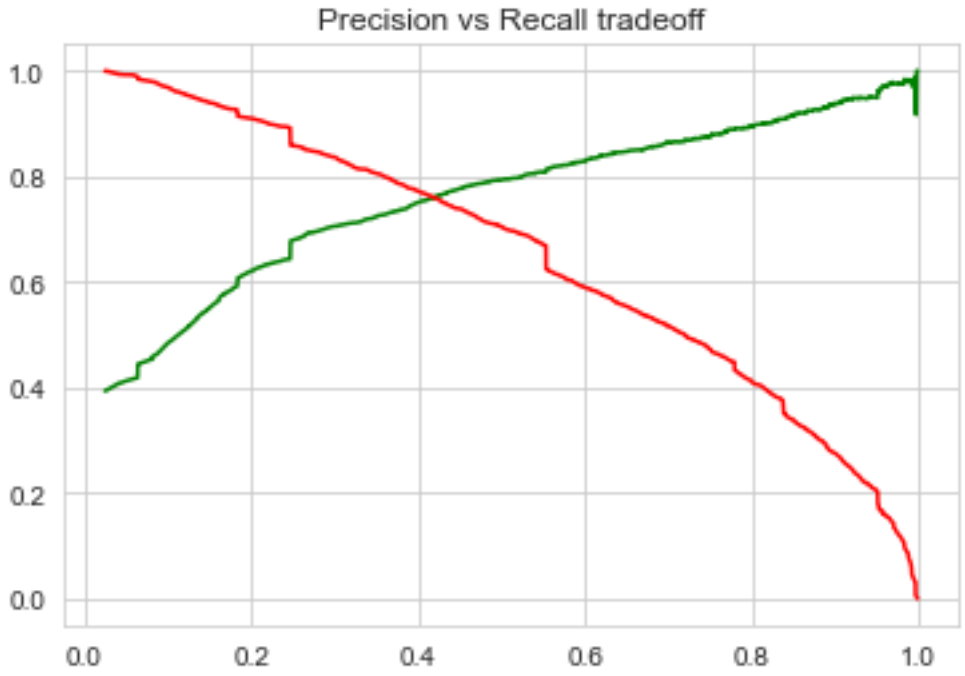




Finding optimal probability cut-off point –

As we can see from the above data we have created points for accuracy , sensitivity and specificity for all probability points from 0 to 0.9.
Out of this we have to choose one as a cutoff point and it is probability cutoff = 0.4 because all the accuracy , sensitivity and specificity are having nearly same value which is an ideal point to consider for as we can't ignore any one from three.

Precision and Recall trade-off



As we can see that there is a trade off between Precision and Recall and the meeting point is nearly at 0.5

Conclusion and Insights -

- The Accuracy, Precision and Recall score we got from test set in aceptable range.
- We have high recall score than precision score which we were exactly looking for.
- In business terms, this model has an ability to adjust with the company’s requirements in coming future.
- This concludes that the model is in stable state.
- Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are :
  - Last Notable Activity_Had a Phone Conversation
  - Lead Origin_Lead Add Form and
  - What is your current occupation_Working Profession