Lead Score Case Study

By

Sakshi Dwivedi

Problem Statement

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is poor around 30%.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- If company get the set of leads which has high probability of conversion, the sales team will be able to focusing more on communicating with the potential leads rather than making calls randomly.

Business Objective:

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

Solution Methodology

- Data cleaning and data manipulation.
    1. Check and handle duplicate data.
    2. Check and handle NA values and missing values.
    3. Drop columns, if it contains large number of missing values and not useful for the analysis.
    4. Imputation of the values, if necessary.
    5. Check and handle outliers in data.

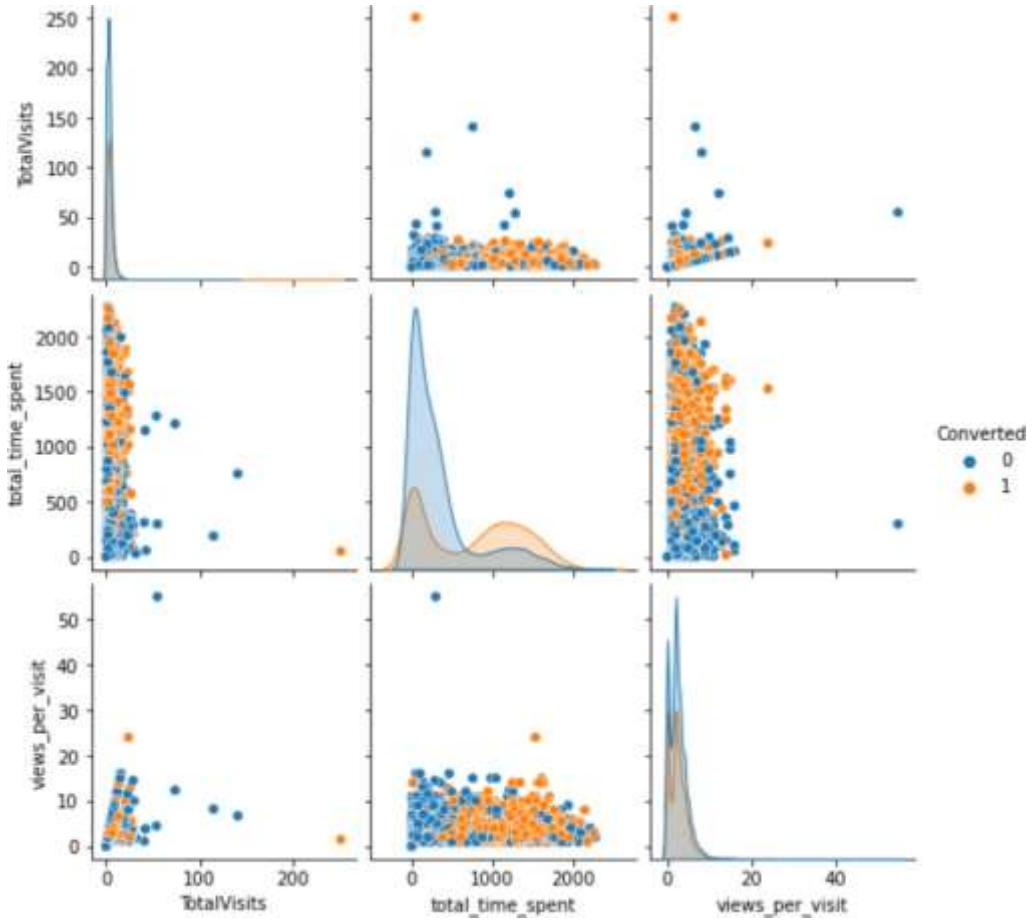
- EDA
    1. Univariate data analysis: value count, distribution of variable etc.
    2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.

- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations

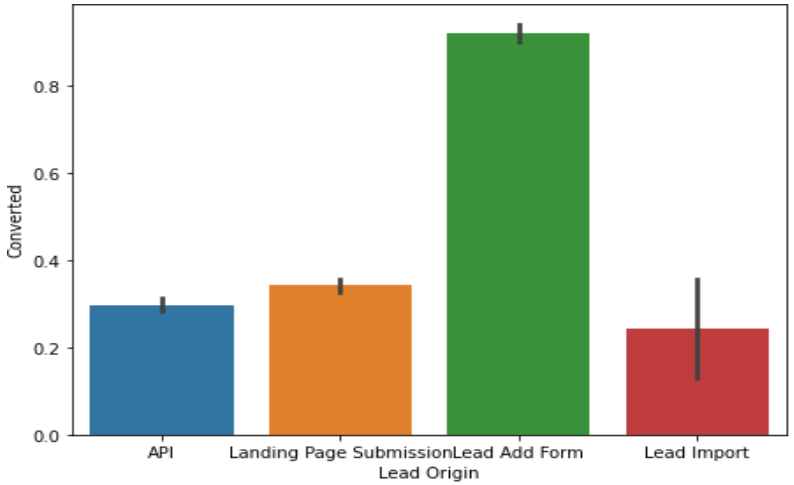
Data Manipulation

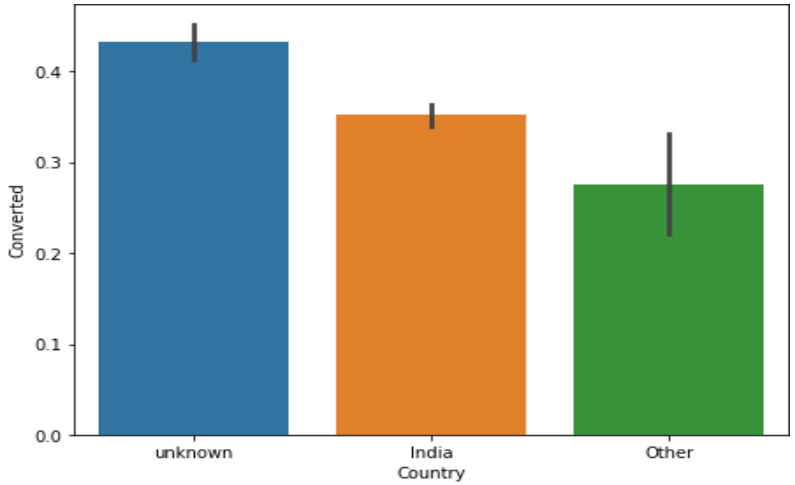
- Input Data have 37 Rows and 9240 columns.
- Dropped 'Prospect ID', 'Lead Number', 'Last Activity’ columns as these don’t have any impact on the target variable.
- Replacing Select value with Null, as both are same.
- Removed the attributes which have more than 30% of missing values.
- Remove all the columns which contains single value throughout the column.
- Replacing Null value with ‘Other’.

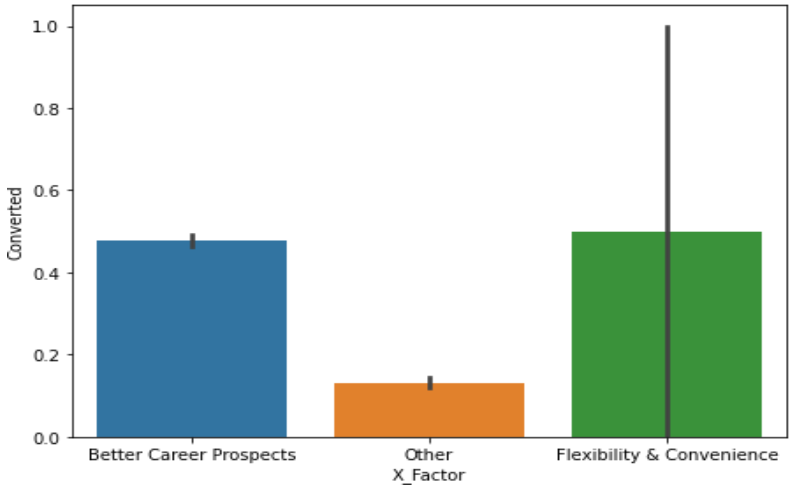
EDA

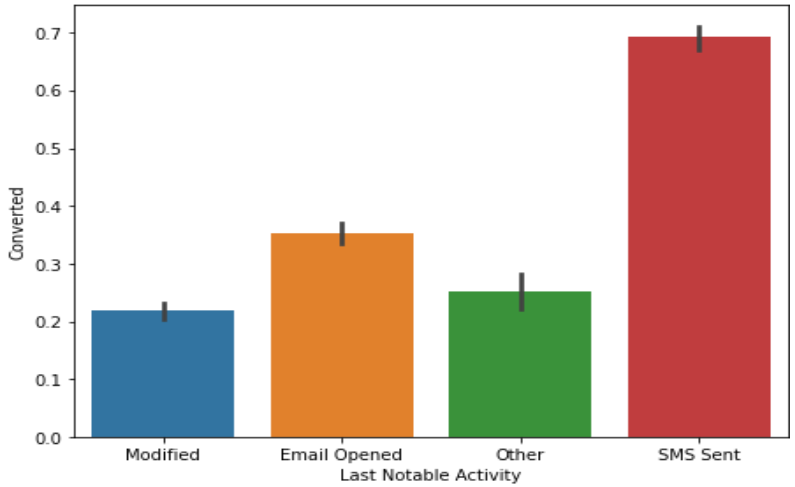


Univariant Analysis









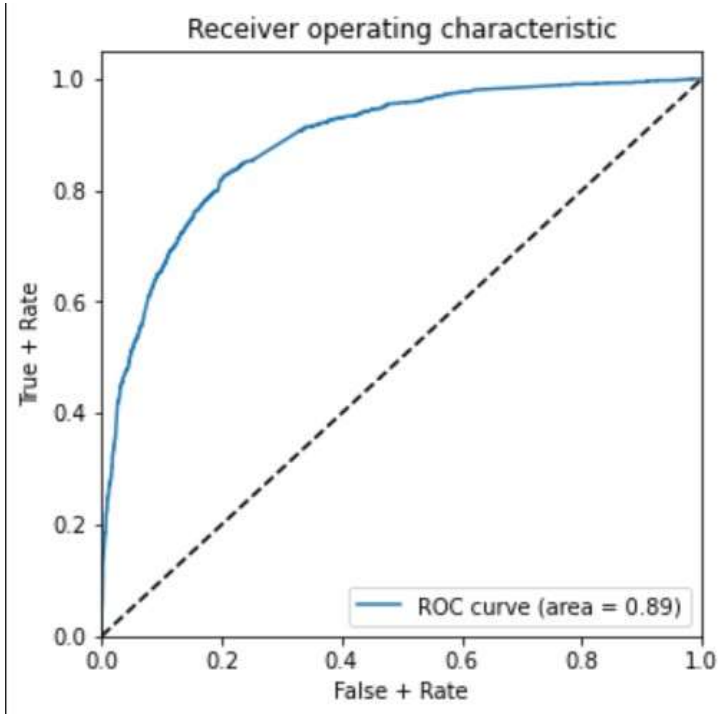
Data Conversion

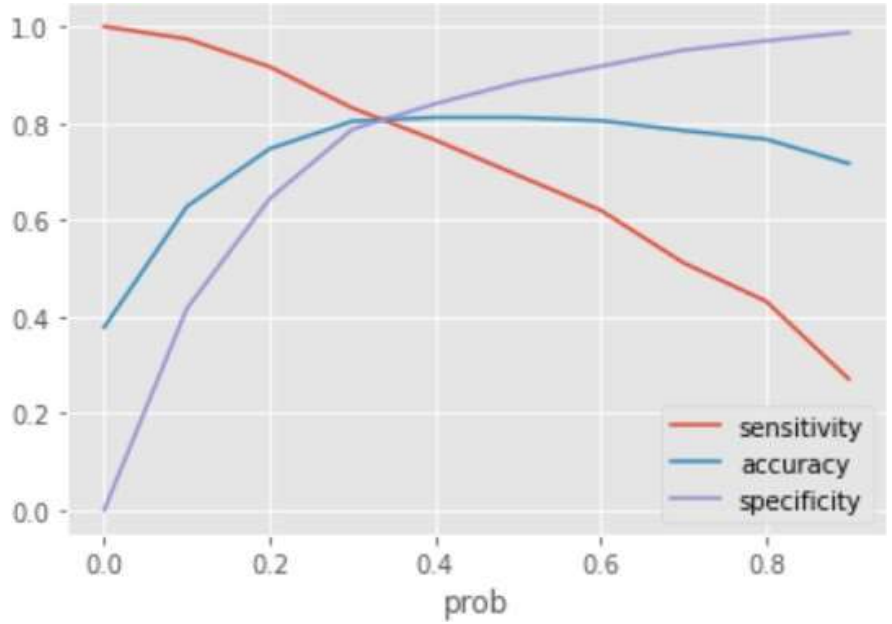
- Numerical Variables are Normalized.
- Dummy Variables are created for object type variables.
- Total Rows for Analysis: 8351.
- Total Columns for Analysis: 29 (including dummy attributes)

Model Building Steps

- After Data cleaning, the dataset is split in the ratio of 70:30 for Train and Test set.
- Used RFE for feature selection and auto feature selection is set to 15.
- Building Model by removing the variable whose p- value is greater than 0.05 and VIF value is greater than 5.
- After manual RFE the features have been reduced to 9.
- Predictions on test data set.
- Validated the accuracy Confusion matrix and classification report.
- Overall accuracy on test data is 82.5%

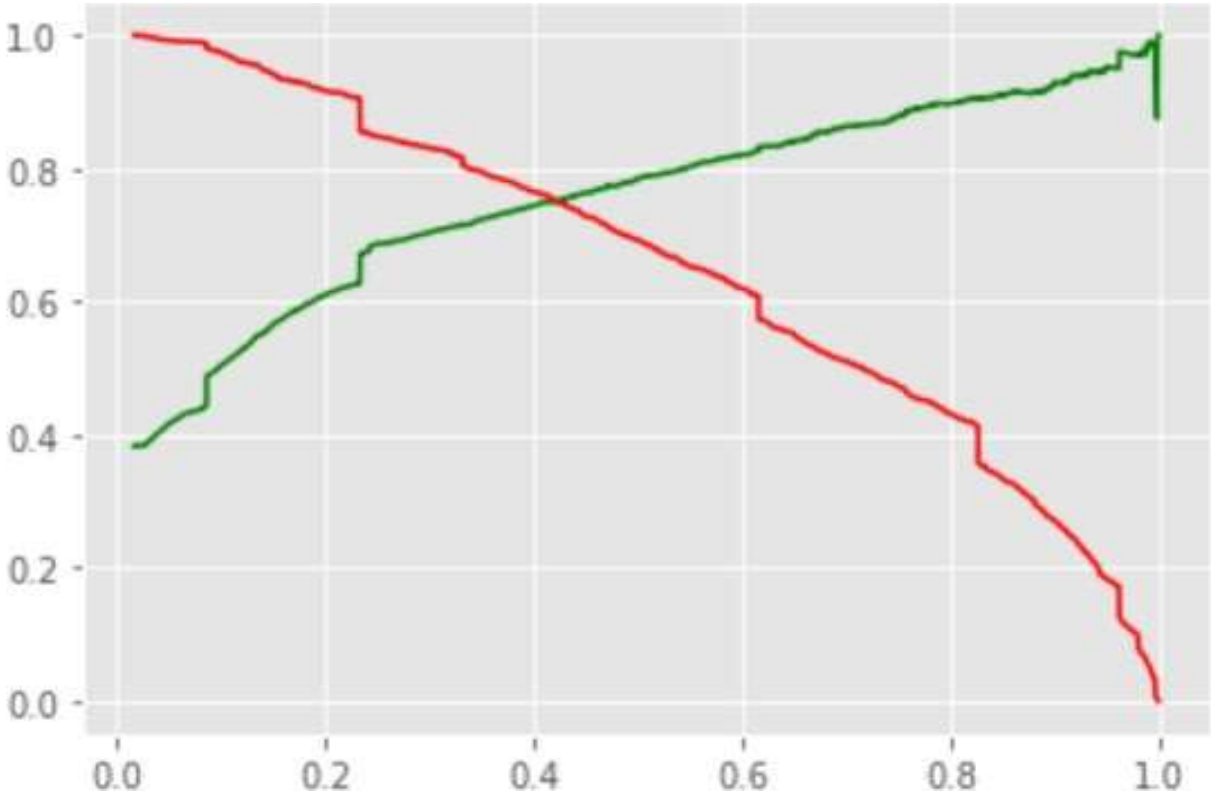
ROC curve and Optimal cutoff





- ROC curve value from above plot is 0.89 which is good.
- From the second graph, we can observe that the cutoff point can be set to 0.34

Precision and Recall Tradeoff



- From the above graph we can conclude that the precision recall trade off point is 0.41

Conclusion

- After the entire exercise it is observed below are top 3 hot leads
  - l'otal l'ime Spent on Website categoíy
  - Lead Oíigin categoíy
  - Total number of visits.
- There is high probability of conversation of leads with the last activity is SMS.
- When the lead origin is Lead add format.
- Time spent on website is also key factor to be considered as the members are interested and have high conversion probability.