Lead Scoring Case Study

Logistic Regression

-by Pradnya R Gangurde

Problem Statement

- X Education is an organization which provides online courses for Industry Professionals. The company marks its courses on several popular websites like Google.
- X Education wants to select most promising leads that can be converted into paying customers.
- Although the company generates a lot of leads only a few are converted into paying customers, wherein the company wants a higher lead conversion. Leads come through numerous modes like email, advertisements on websites, google searches , etc.
- The company has had 30% conversion rate through the whole process of turning leads into customers by approaching those leads which are to be found having interest in taking the course. The implementation process of the lead generating attributes are not efficient in helping conversions.

Business Goals

- The company requires a model to be built for selecting most promising leads.
- Lead score to be given to each leads such that it indicates how promising the lead could be. The higher the lead score the more promising the lead to get converted, the lower it is the lesser the chances of conversion.
- The model to be built In Lead conversion Rate around 80% or more.

Strategy

- Import Data
- Clean and prepare the acquired data for further analysis.
- Exploratory Data Analysis for figuring out most helpful attributes for conversion.
- Scaling Features
- Prepare the data for Model Building.
- Assign a lead score for each lead.
- Test the Model on the Train set
- Evaluated model by different measures and metrics
- Test the model on Test set
- Measure the accuracy of the model and other metrics for evaluation.

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 amount of missing values and not useful for the analysis.
4. Imputation of the values, if necessary.
5. Check and handle outliers in data.

- EDA

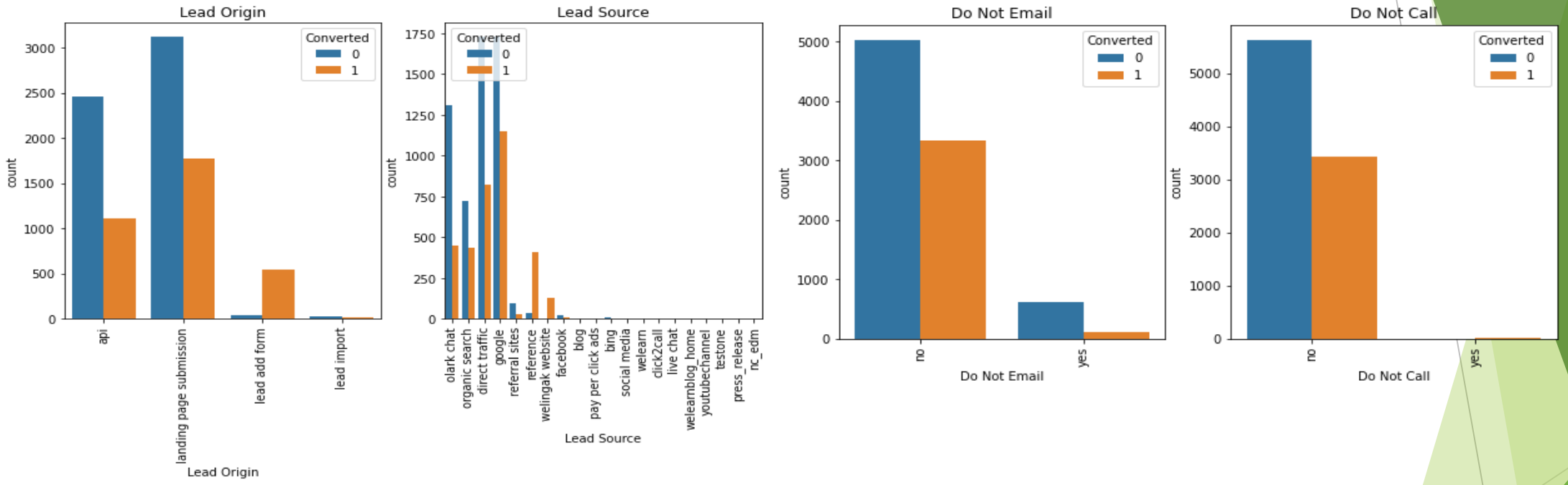
1. 1. Univariate data analysis: value count, distribution of variable etc.
2. 2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.

- 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

- Total Number of Rows =37, Total Number of Columns =9240.
- Single value features like “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” etc. have been dropped.
- Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.
- Dropping the columns having more than 35% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’.

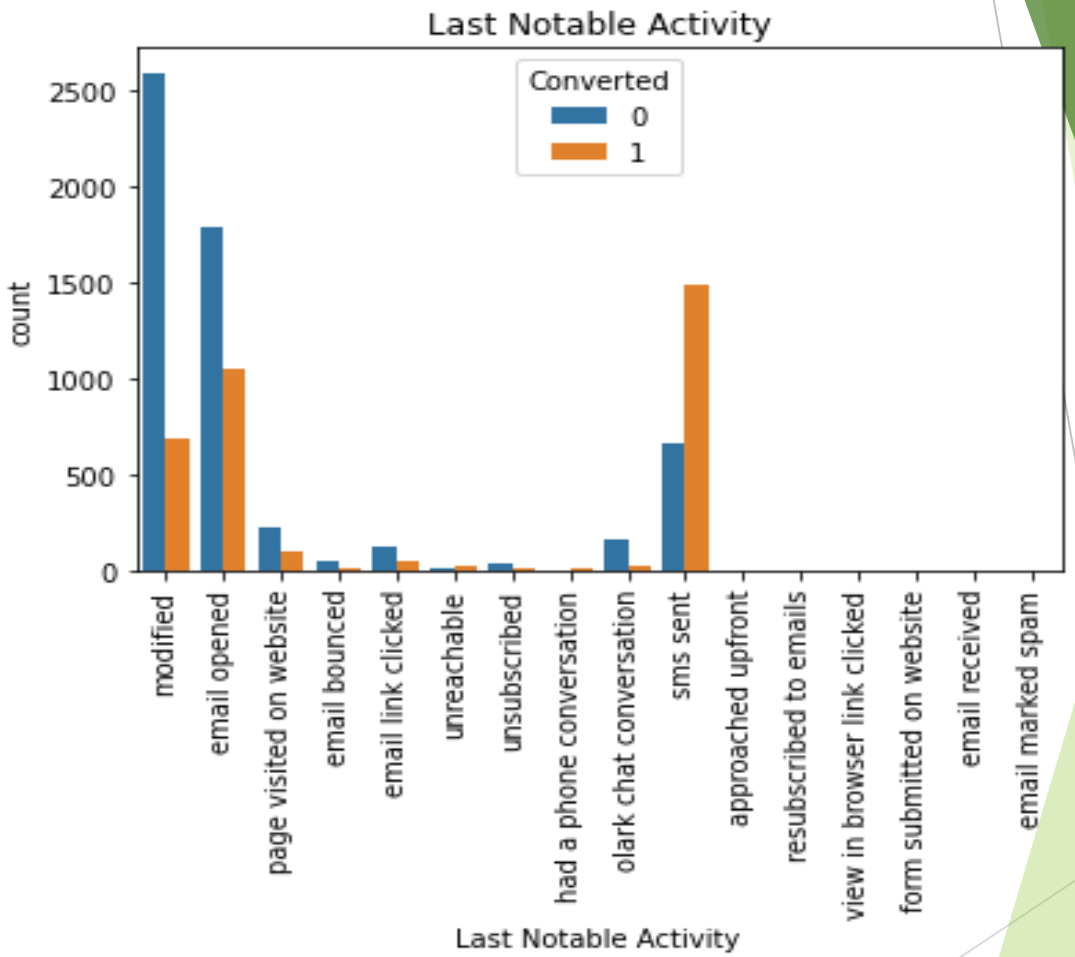
Exploratory Data Analysis



Lead source vs Converted : Google searches has high conversions compared to other modes, whilst reference has had high conversion rate.
Lead Origin vs Converted : Landing Page submission has high conversions compared to other modes, whilst reference has had high conversion rate.
Do Not Email/call Vs Converted : Google searches has high conversions compared to other modes, whilst reference has had high conversion rate.

Last Notable activity Vs Converted

Last notable activity i.e. modified has high conversions compared to other modes, whilst reference has had high conversion rate.



Data Conversion

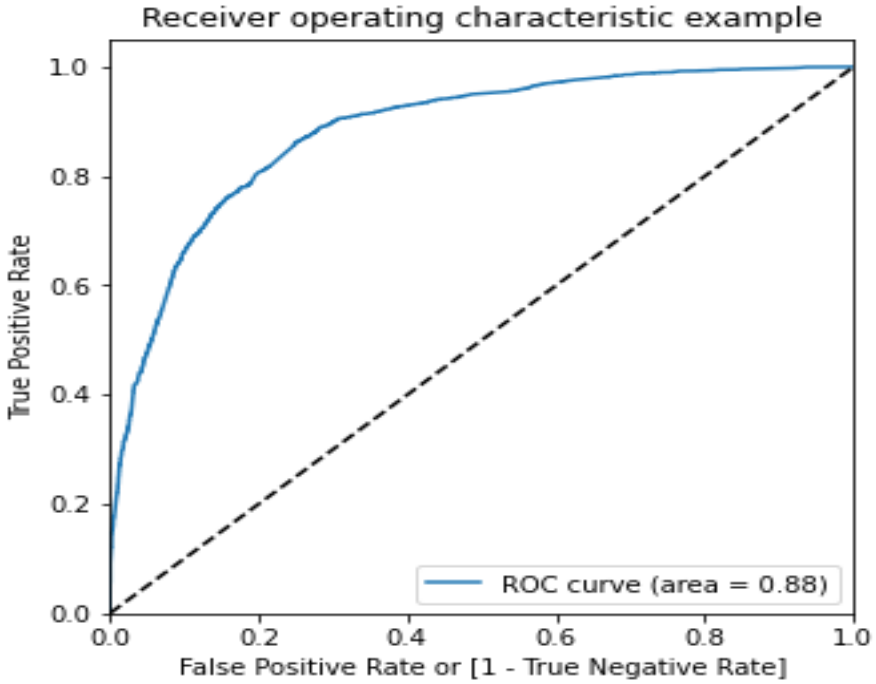
- Numerical Variables are Normalised
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 8792
- Total Columns for Analysis: 43

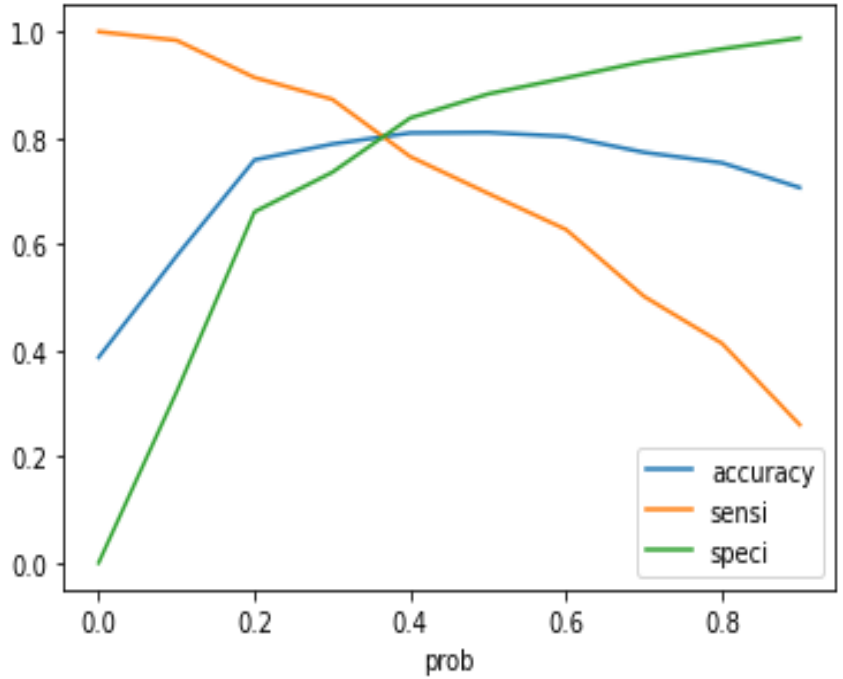
Model Building

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 than 0.05 and VIF value is greater than 5
- Predictions on test data set
- Overall accuracy 81%

ROC Curve





* Finding Optimal Cut off Point:

1. Optimal cut off probability is that probability where we get balanced sensitivity and specificity.
2. From the second graph it is visible that the optimal cut off is at 0.35.

Conclusion

It was found that the variables that mattered the most in the potential buyers are (In descending order) :

- The total time spend on the Website.
- Total number of visits.
- When the lead source was:

a. Google
b. Direct traffic
c. Organic search
d. Welingak website

- When the last activity was:

a. SMS
b. b. Olark chat conversation

- When the lead origin is Lead add format.
- When their current occupation is as a working professional.

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.