



Lead Scoring Case Study

Group Members:

Kaushal Nagrecha

Anant Sawant

Abhinav Pandey



Problem Statement

- An education company named *X Education* sells online courses to industry professionals.
- Once these people land on the website, they might browse the courses or fill up a form for the course. When these people fill up a form providing their email address or phone number, they are classified to be a lead.
- The typical lead conversion rate at *X Education* is around 30%.
- To make the *lead conversion process* more efficient, the company wishes to identify the most potential leads, also known as *Hot Leads*.



Goals of the Case Study

- 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. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- There are some more problems presented by the company which our model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well.




Business Objectives

- 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.


Approach



- 
- Data cleaning and data manipulation -
 - Check and handle duplicate data.
 - Check and handle NA values and missing values.
 - Drop columns, if it contains large amount of missing values and not useful for the analysis.
 - Imputation of the values, if necessary.
 - Check and handle outliers in data.
 - EDA -
 - Univariate data analysis: value count, distribution of variable etc.
 - 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.

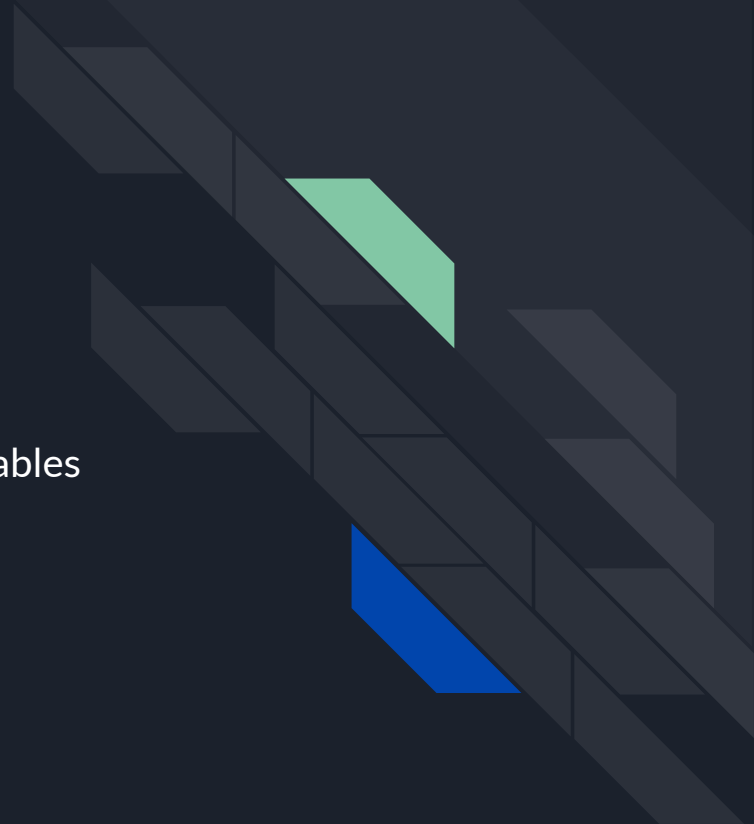
Data Manipulation



- 
- Total Number of Columns =37, Total Number of Rows =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’.


Data Conversion

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

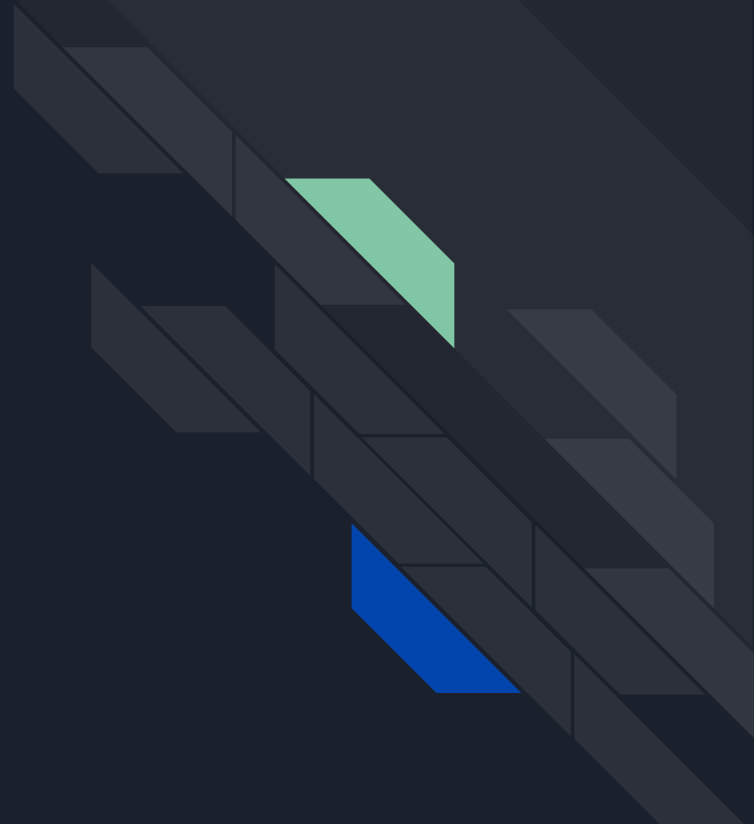


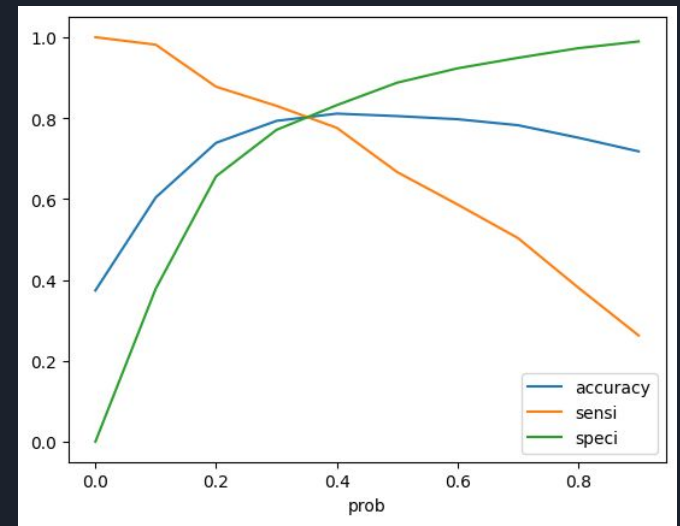
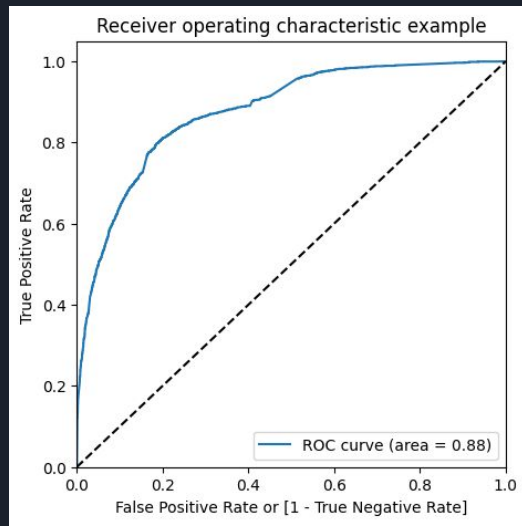
Model Building



- 
- 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






- 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





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

1. The total time spent on the Website.
2. Total number of visits.
3. When the lead source was:
 - a. Google
 - b. Direct traffic
 - c. Organic search
 - d. Welingak website
4. When the last activity was:
 - a. SMS
 - b. Olark chat conversation
5. When the lead origin is Lead add format.
6. When their current occupation is as a working professional.

Keeping these in mind X Education can identify the `Hot Leads` and maximize their conversion rate.