# Lead Scoring Case Study

Batch IIIT-B DS 53

# X- Education

**PROBLEM STATEMENT**

- X Education sells online courses to industry professionals.
- X Education generates plenty of leads. However, its lead conversion rate is very poor. Currently if they acquire 100 leads, only about 30 of them are converted.
- As of now the company's process includes reaching out to all the 100 leads. This is very inefficient and time-consuming. The company wishes to use machine learning techniques to identify potential leads (Hot Leads) thereby increasing lead conversion rate, saving costs and improving efficiency.

**BUSINESS OBJECTIVE**

- **Intelligent Lead Filtering:** Create a machine learning model to identify promising leads.
- **Increase in conversion rate:** The current conversion rate is ~ 30%. They want the model to help increase this conversion rate to ~ 80%.
- **Efficient Lead Management:** Assign a lead score that would help sales follow up with the genuinely interested customer and thereby increase the overall potential revenue.

# Solution Methodology

**WE HAVE USED THE TRADITIONAL DATA MODELLING FLOW THAT STARTS FROM CLEANING THE DATA, TO MODELLING THE DATA USING TRAINING DATA AND THEN TESTING THE SAME TO VALIDATE OUR FINDINGS.**

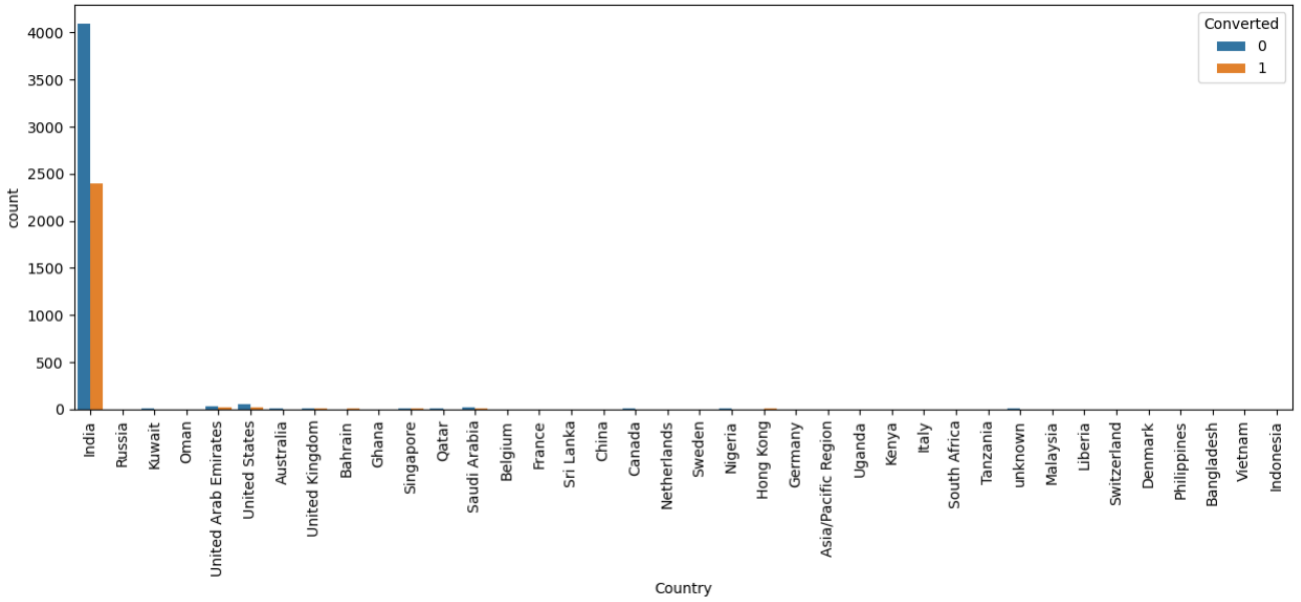
- **Data cleaning and data manipulation.**
    1. Check and handle duplicate data.
    2. Check and handle NA values and missing values.
    3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
    4. Imputation of the values, if necessary.
    5. Check and handle outliers in data.
- **EDA**
    1. Univariate data analysis: value count, distribution of variable etc.
    2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.
- **Model Building using RFE:**
    - Feature Scaling & Dummy Variables and encoding of the data.
    - Classification technique: logistic regression used for the model making and prediction.
    - Validation of the model.
- **Conclusions and recommendations.**

# DATA CLEANING AND MANIPULATION (1/2)

## DE-DUPLICATION OF DATA AND HANDLING OF MISSING VALUES AND OUTLIERS.

- We started the analysis with a dataframe of 9240 rows and 37 columns.
- There was a mix of categorical and numeric variables. For e.g. Search, Magazine, etc that had just “Yes” and “No” values, Total Visits, Total time, etc that were continuous numeric variables.
- We started off by deleting the Prospect ID and Lead Number columns as they were just identifiers and were not really adding value to the model.

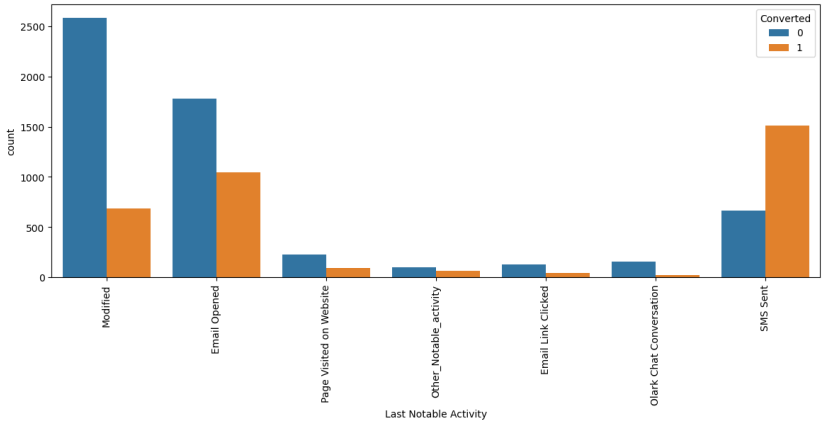
- We then deleted the columns that had more than 45% missing values as substitution of data at that large scale would affect the data sanctity.
- For the remaining columns with missing values, we plotted count plots to understand the distribution of data against conversions and accordingly substituted values. For eg, in the case of country we got the plot as you see on the right. Basis this plot we decided to substitute NaN values with “India” in country column.

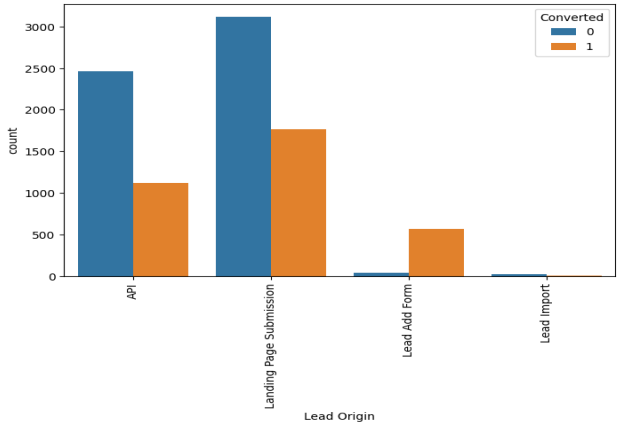


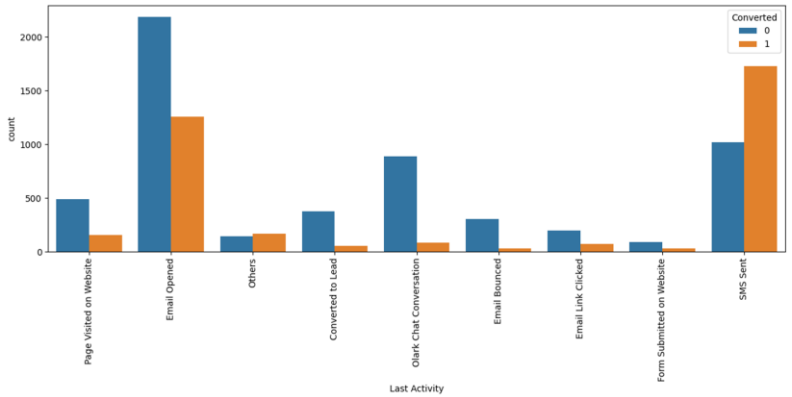
# DATA CLEANING AND MANIPULATION(2/2)

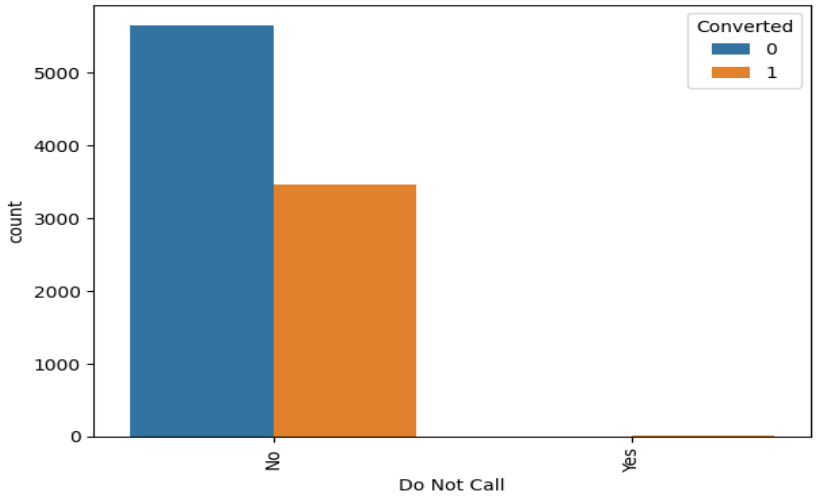
## DE-DUPLICATION OF DATA AND HANDLING OF MISSING VALUES AND OUTLIERS.

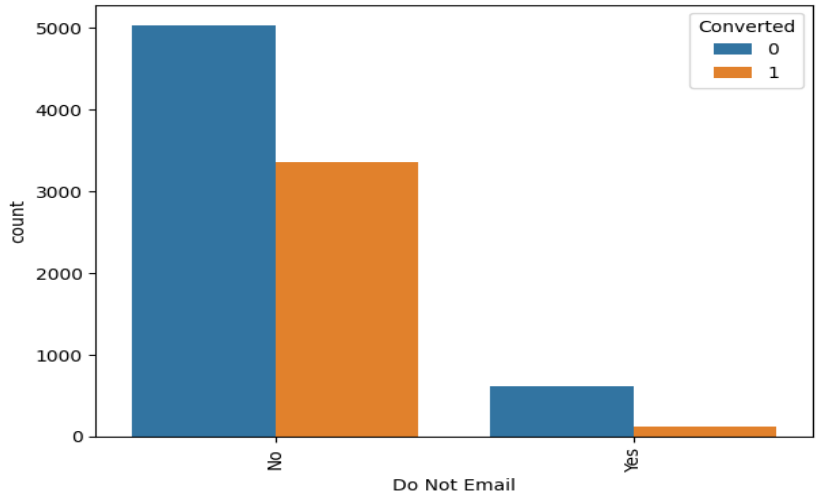
- Here are some of the other values that we substituted using the same approach.



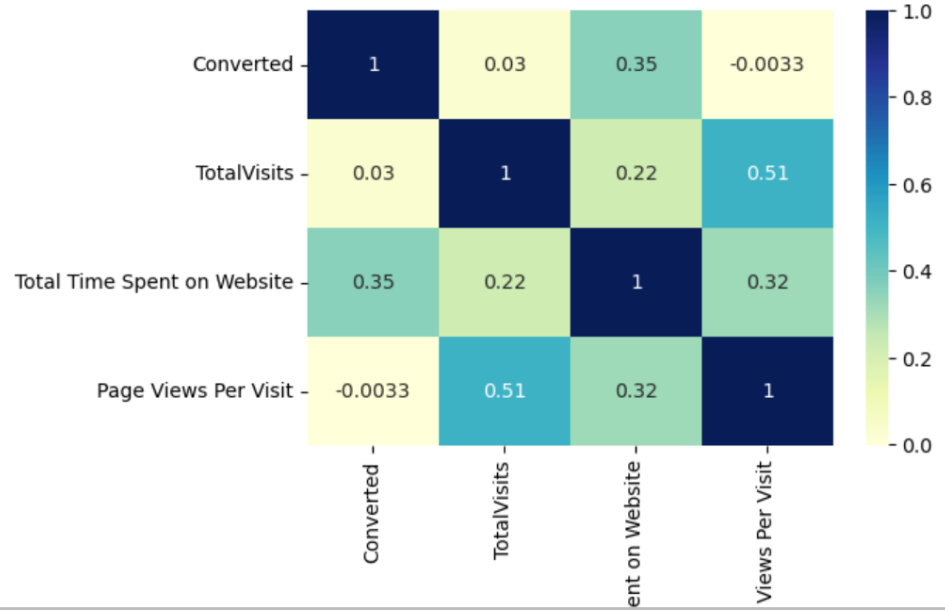



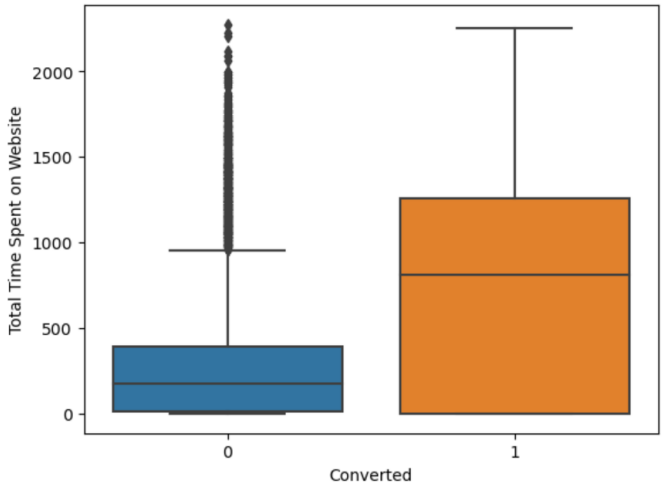



# EXPLORATORY DATA ANALYSIS

## UNIVARIATE AND BIVARIATE ANALYSIS TO UNDERSTAND IF THERE ARE ANY CORRELATIONS BETWEEN VARIABLES.

- Then we analyzed different variables. Here are some of the inferences:
  - API and Landing Page Submission bring higher number of leads as well as conversion.
  - Lead Add Form has a very high conversion rate but count of leads are not very high.
  - Lead Import and Quick Add Form get very few leads.
  - Maximum number of leads are generated by Google and Direct traffic.
  - Conversion Rate of reference leads and leads through welingak website is high.
  - To improve overall lead conversion rate, focus should be on improving lead converion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.
- We then checked for bi-variate correlations to see if there are any numeric columns that we need to deal with
- We also handled outliers by calculating the percentile values for a few features and removing the values less than the 1 and 99 percentiles.

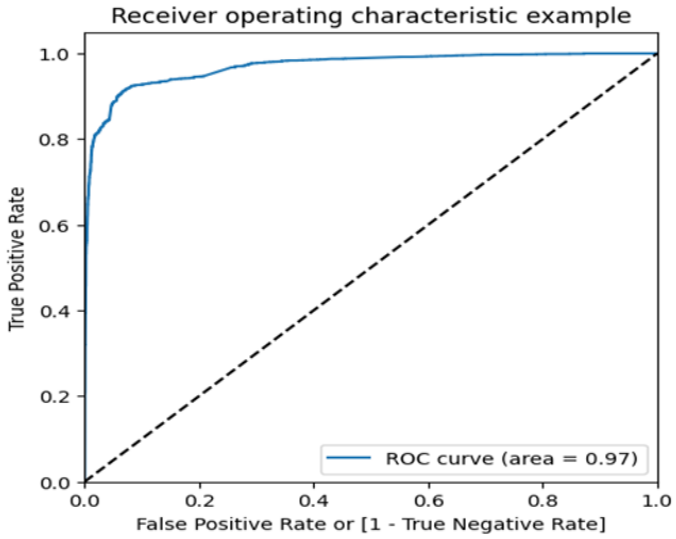


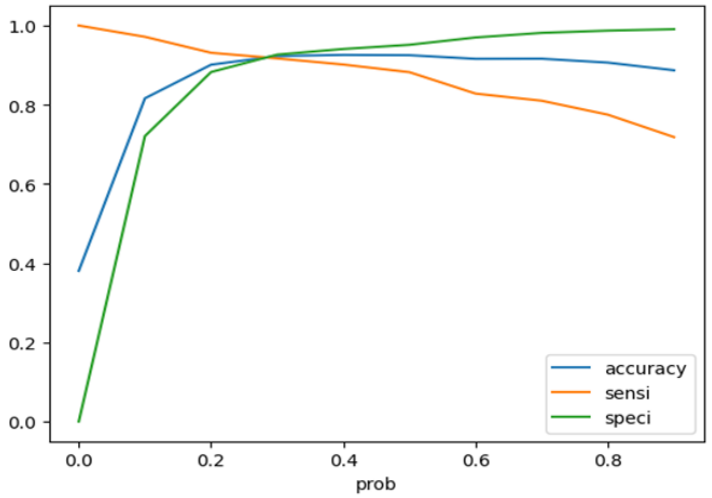



# MODEL BUILDING

## DUMMY FEATURES, SCALING, SPLITTING DATA FOR TRAIN AND TEST, MODEL BUILDING AND VALIDATION.

- We then started creating dummy variables for the categorical variable. We ended up having 56 columns when these dummy variables were created. We then used RFE to select the top 15 features that were relevant for the model.
- The 15 features we got are: *'Total Time Spent on Website', 'Lead Origin_Lead Add Form','Lead Source_Direct Traffic', 'Lead Source_Referral Sites', 'Lead Source_Welingak Website', 'Last Activity_SMS Sent', 'Last Notable Activity_Modified','Last Notable Activity_Olark Chat Conversation', 'Last Notable Activity_SMS Sent', 'Tags_Closed by Horizzon', 'Tags_Interested in other courses', 'Tags_Lost to EINS', 'Tags_Other_Tags', 'Tags_Ringing', 'Tags_Will revert after reading the email'*
- We then created 3 iterations of the model by checking for p values, VIF values and we reached a model with decent accuracy, sensitivity metrics.
- We then plotted the ROC curve and then the probability distribution curve to find the probability cutoffs (0.35).



# CONCLUSIONS AND RECOMMENDATIONS

## WHAT WE FOUND AND WOULD LIKE TO SUGGEST TO X-EDUCATION

- It was found that the variables that mattered the most in the potential buyers are:
  - The total time spend on the Website.
  - Total number of visits.
- When the lead source was:
  - Google
  - Direct traffic
- When the last activity was:
  - SMS Sent
  - Email opened
  - Olark Chat
- When the lead origin is Lead add format.
- When their current occupation is as a working professional.

***BY FOCUSSING ON USERS WITH THE ABOVE ATTRIBUTES AND CALLING AND FOLLOWING UP WITH THEM, X-EDUCATION HAS A HIGH CHANCE OF INCREASING CONVERSIONS AND ACHIEVING BETTER OPERATIONAL EFFICIENCTY.***