# Lead scoring case study

Aishwarya Pradeep & Balaji Sriram

# Problem Statement

- X Education, an education company sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. Their typical lead conversion rate is around 30%, which is very poor.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

# Overall approach

- Analyse the “Leads.csv” data and carried out the following steps:
  - Check for missing values and take necessary actions:
    - Delete if the percentage of null values are greater than 40%
    - Check if the column values have “Select” and convert into NaN
    - Delete if one of the category value is skewed.
  - Treating null values:
    - Deleting or Imputing columns which has null values.
- Data Imbalance analysis and Dummy variables creation
- Split the overall data into Train and Test data
- Data Modeling and evaluation

# Dataset details

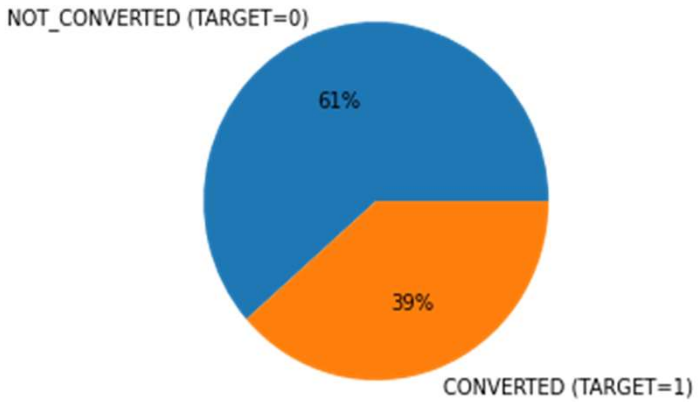
- “Leads.csv” dataset contained 9240 rows and 37 columns.

- Upon data understanding, it was found that:
  - 7 columns had null values greater than 40% and those were dropped
  - 3 columns where were unnecessary (based on data dictionary)
  - 12 columns had either only one category or two category values (which are >99%)  were dropped

- 15 relevant features were considered for model building and evaluation.

# Data Imbalance check

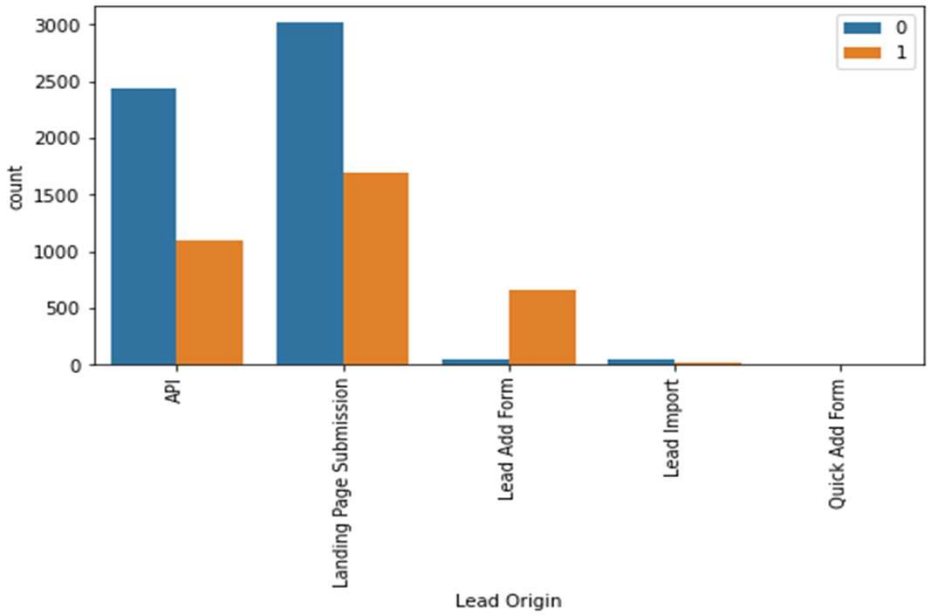
- 61% of the leads were unsuccessful (i.e. leads not converted)
- 39% of the leads converted successfully.
- The target variable is not skewed and the data looks balanced



# Analysis of Categorical variables VS Target variable

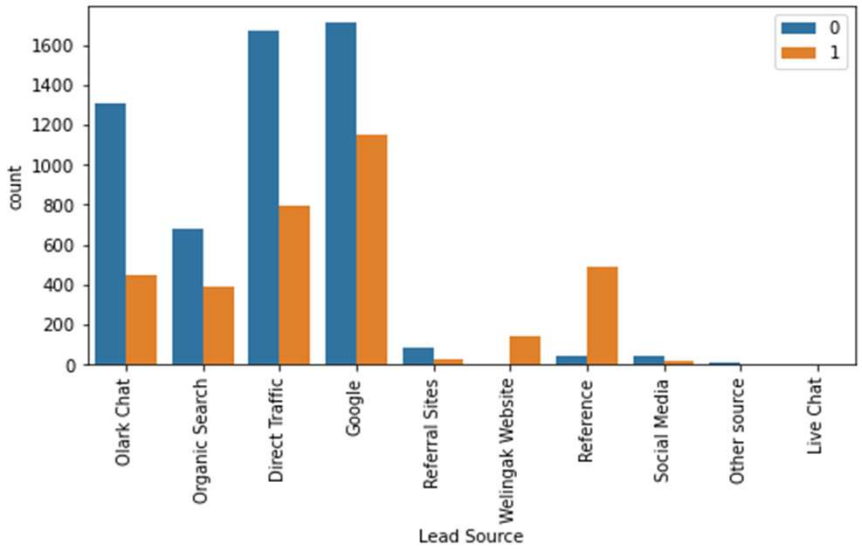
# Lead Origin VS Converted

- Most leads were originated from Landing page submission followed by API and their conversion rate seems good.
- Conversion rate on Lead add form is higher than API & landing page submission but the count is very less.
- If Lead source is Add Form, the ratio of lead conversion is very high.



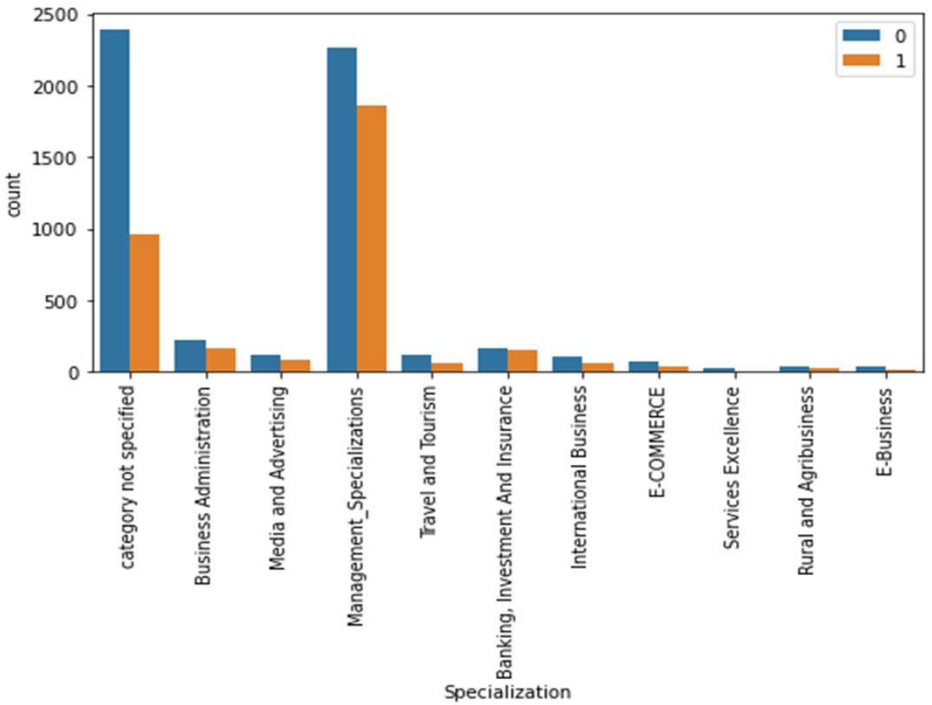
# Lead Source VS Converted

- Maximum lead sources are generated via Google followed by direct traffic.
- Conversion Rate of reference leads and leads through Welingak website is high.



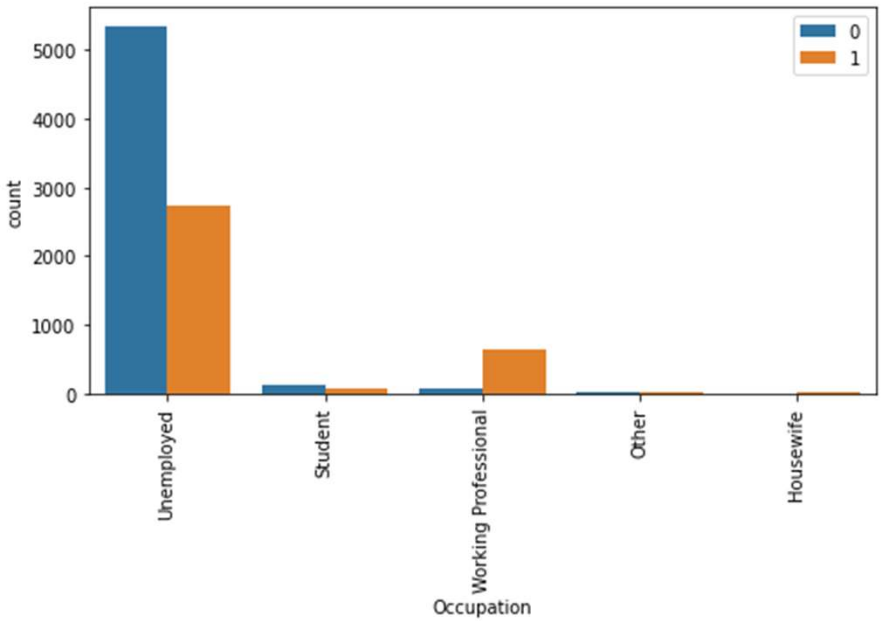
# Specialization VS Converted

- Leads with Management specialization has the highest conversion rate



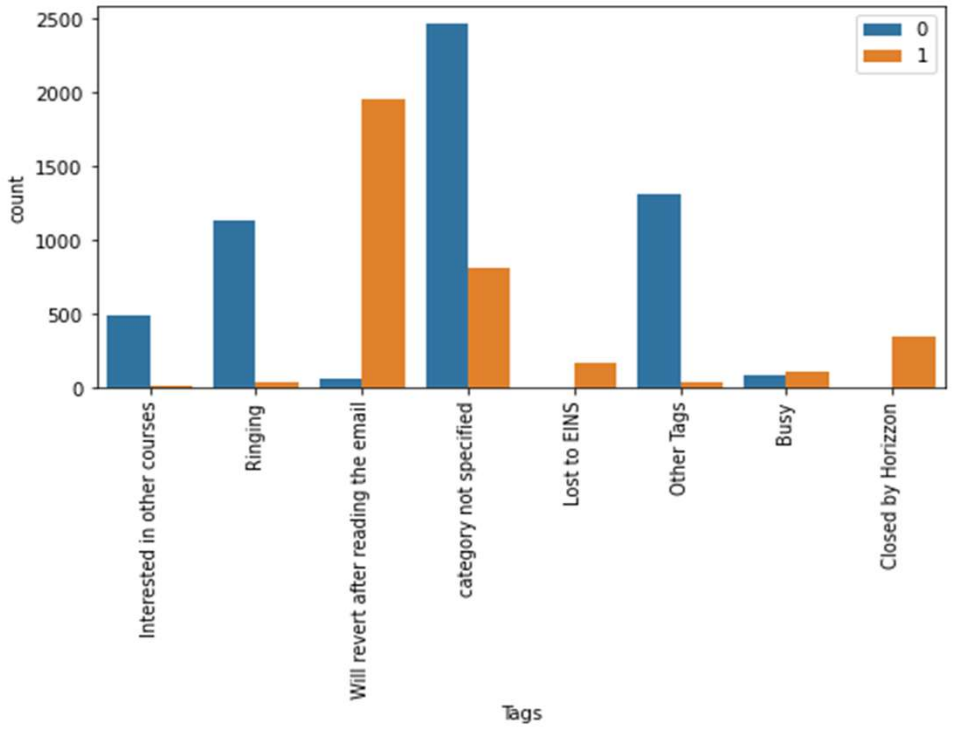
# Occupation VS Converted

- Customers who are Unemployed have a high chances of leads conversion.
- Working professional opting for the course have high probability of enrolling.

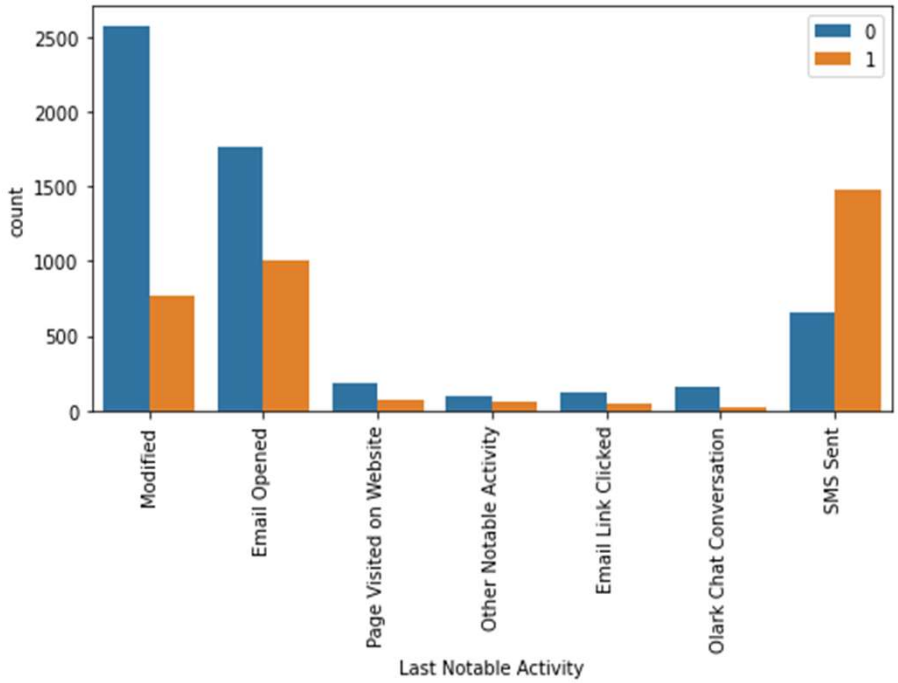


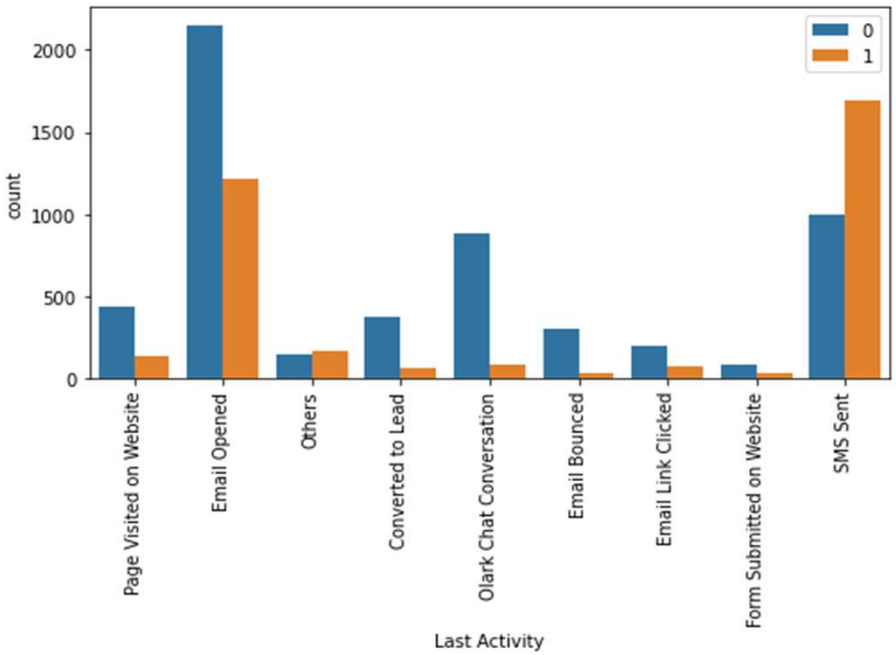
# Tags VS Converted

- The current status of the lead "Will revert after reading the email" has the maximum probability of lead conversion followed by "Closed by Horizon"



# Last Notable Activity & Last Activity VS Converted



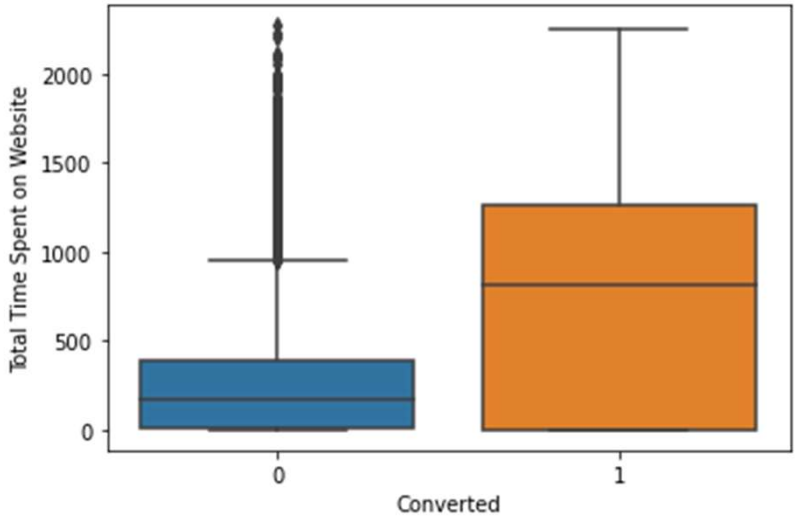




- After knowing about the course, the next actions would be sending SMS and emails which constitute the maximum lead conversion

# Analysis of Continuous variable VS Target variable

# Total Time Spent on Website VS Converted

- The conversion rate for total time spent on the website is more (i.e. Leads spending more time on the website are more likely to be converted.)

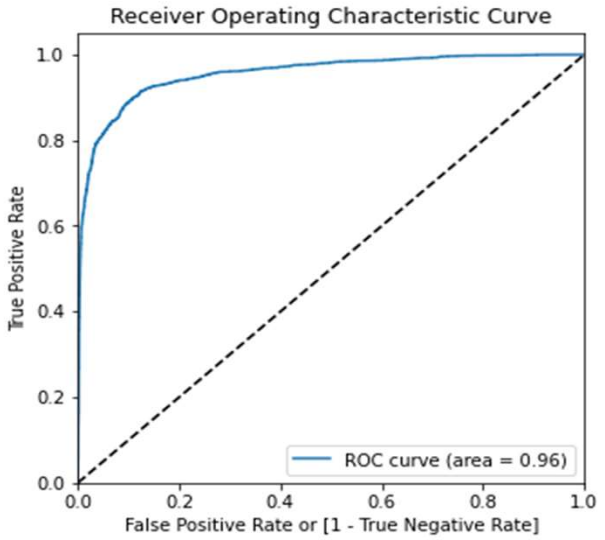


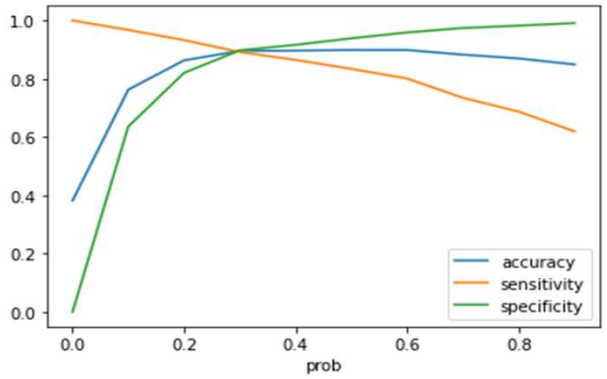
# Data Pre-processing

- Outlier treatment:
  - Total Visits and Page Views Per Visit had some outliers.
  - Based on the understanding of 50, 75, 90, 95 and 99 percentile, outlier treatment was carried out by imputing with median.
- Mapping Binary data:
  - ‘Do not email’ and ‘A free copy of mastering the interview’ contains value in Yes/No format which is mapped to 1/0.
- Dummy variable creation:
  - Dummy variables were created for ‘Lead Origin', 'Occupation', 'Last Notable Activity', 'Specialization', 'Lead Source', 'Tags', 'City', 'Last Activity’ which will be useful for model building.
- Train-Test split
  - It’s of 70% (train data)-30% (test data)
  - Random state = 100
- Scaling:
  - Y will only have the target variable (i.e., Converted) and x will have all the variables apart from target variable.
  - Technique: Standard Scaler

# Data Modeling

- Feature selection using RFE
- Assessing the models using statsmodel
  - Making a VIF dataframe for all the variables
- Predictions on the Train data set
  - Create a new column ‘Predicted’ if probability > 0.5
  - Dropping columns which has p-value > 0.5
- Calculating Accuracy, Confusion matrix on the Train data set
  - Accuracy = 89.54%
  - Sensitivity = 89.16%
  - Specificity = 89.77%
- Plotting the ROC curve
- Finding the optimal cut-off point of 0.3
- Precision = 84.32%
- Recall = 89.16%



# Data Modeling, cont’d..

- Predictions on the Test data set
  - Scaling the continuous variables
- Calculating Accuracy, Confusion matrix on the Test data set
  - Accuracy = 90.31%
  - Sensitivity = 89.47%
  - Specificity = 90.85%

# Conclusion on Hot Leads

- Focus can be given more to the following customers who can be converted to Hot leads:

  - Customers who spend significant time on the website
  - Working Professionals & Unemployed
  - Customer's last activity is SMS sent / Email opened

| Parameters | Train Data | Test Data |
|---|---|---|
| Accuracy | 89.54% | 90.31% |
| Sensitivity | 89.16% | 89.47% |
| Specificity | 89.77% | 90.85% |

- From the above, we can conclude that the model seems to be performing well.