LEAD SCORING CASE STUDY

ARUNDHATI DESHPANDE

IMRAN TOTTI SHAIK

BUSINESS UNDERSTANDING

- Business:
  - An education company named X Education sells online courses to industry professionals and markets its courses on several websites and search engines like Google.
  - Leads are those professionals with interest in courses, land on company website and browse for courses or fill up a form for the course or watch some videos. Moreover, the company also gets leads through past referrals.
  - Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not.
- Business Problem:
  - The typical lead conversion rate at X education is around 30%. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
  - If they successfully identify hot 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.
- Business Solution:
  - X Education wants to select the most promising leads, i.e., the leads that are most likely to convert into paying customers.

DATA UNDERSTANDING

- Data Problem:
  - A data set with around 9000 past leads has been provided along with data dictionary.
  - This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc.
  - These attributes may or may not be useful in ultimately deciding whether a lead will be converted or not.
  - The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted.
- Data Solution:
  - 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.
  - Make sure this model helps in meeting the CEO’s target of lead conversion rate to be around 80%.

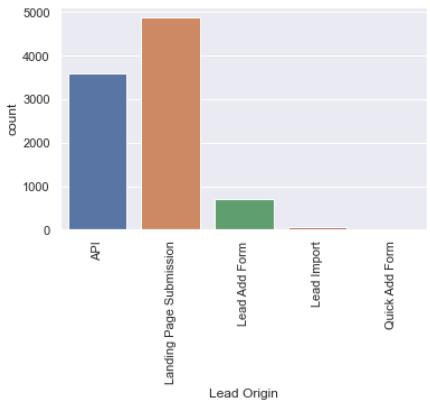
Problem Solving Approach Followed

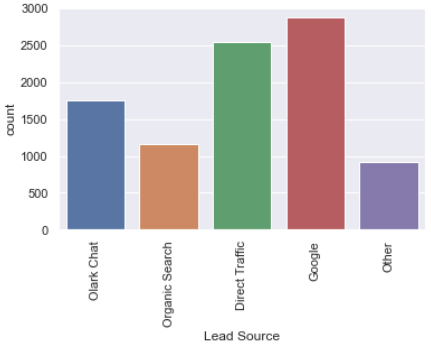
1. Data Cleaning:
   - Dropped the columns which have data entered manually by the sales team and the columns with 40% of its members as missing values.
   - Handled columns with nulls and select as their members using imputation.
   - Dropped highly skewed categorical columns.
   - Performed Outlier treatment using box plots on numerical columns.
2. Exploratory Data Analysis: Analysed trends in the dataset using univariate, bivariate and multivariate analysis by plotting different charts.
3. Data Preparation:
   - Dummy variable creation for categorical variables
   - Standardization of the scales of continuous variables
   - Test-train split of the data
4. Model Building: Built classification model to predict the hot leads.
5. Model Evaluation: Validated the model.
6. Conclusion or Inferences: Provided insights using all the above steps.

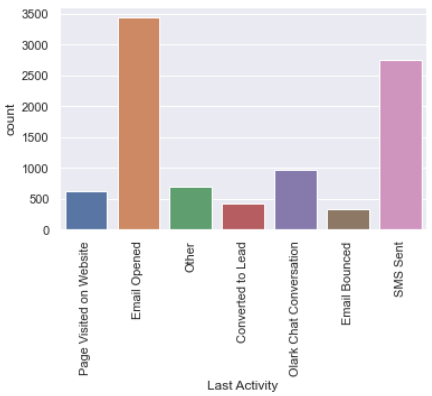
Insights from EDA

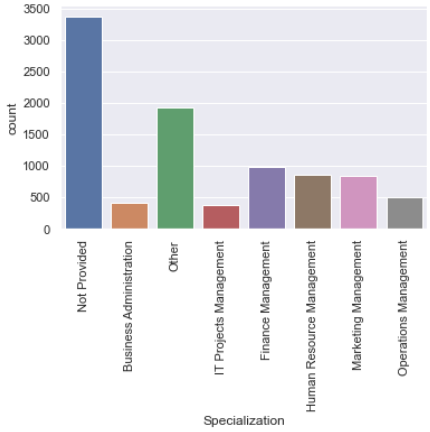
Univariate Analysis:

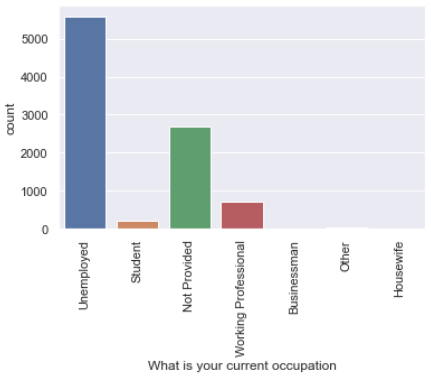
- In Lead Origin column, frequently occurring values are API and Landing page submission
- In Lead Source column, Direct Traffic and Google are the two main source for Leads

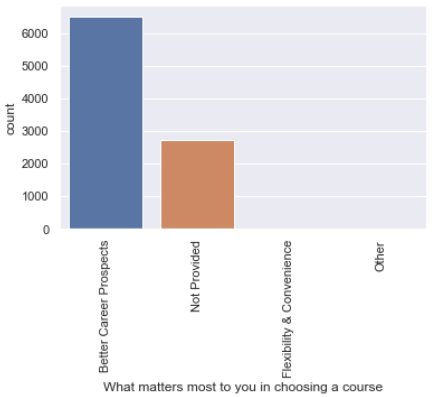


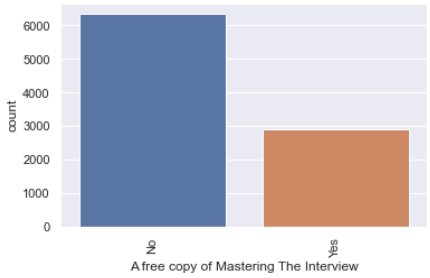


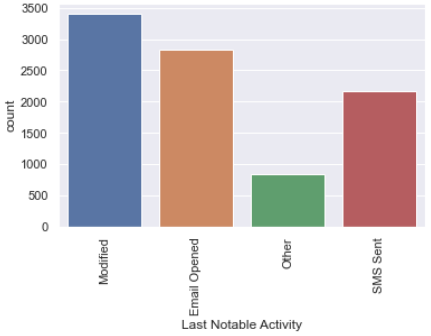








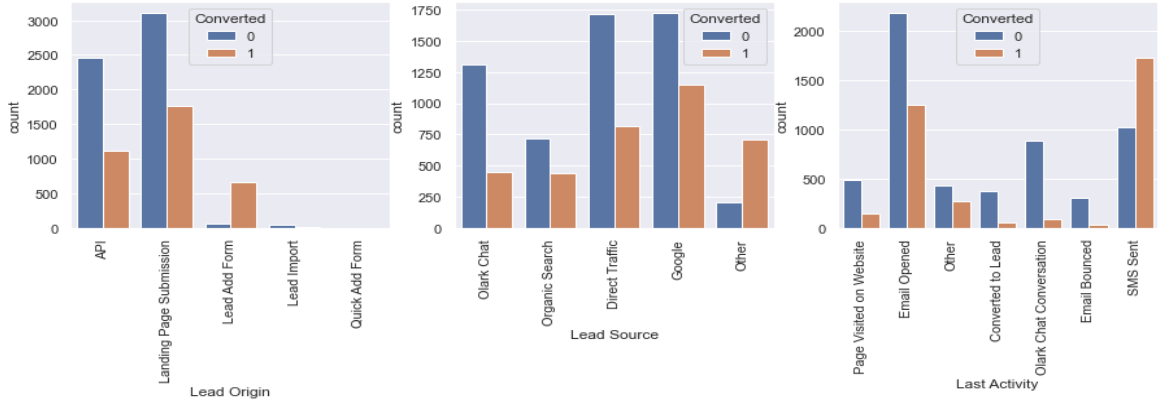


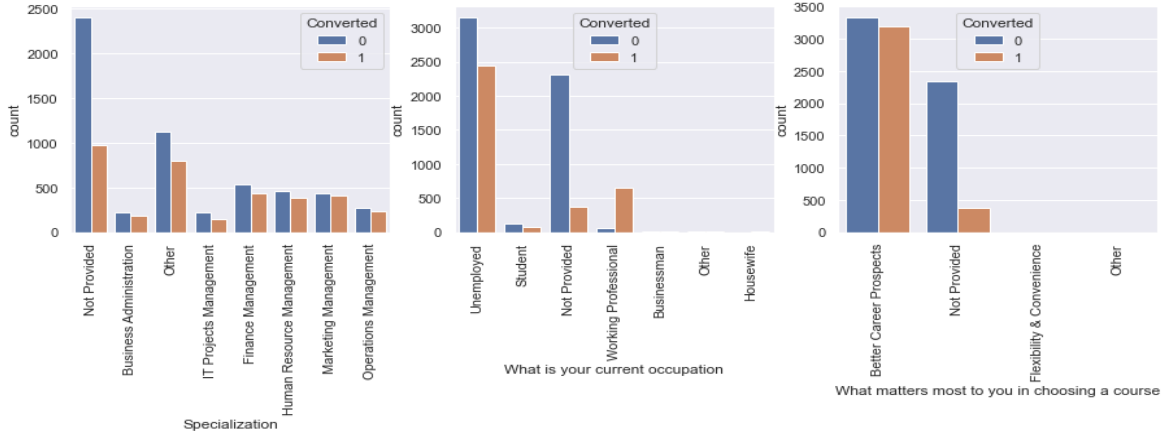


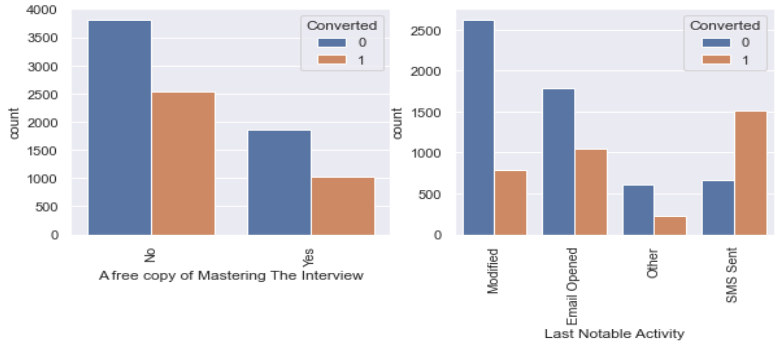
Insights from EDA

Bivariate Analysis:

- Rate of Conversion is very high with Working professionals.
- Rate of Conversion is less in leads originated from Landing page submission and high through lead add form.
- Rate of Conversion is less in leads sourced from Direct Traffic and Olark chat.
- In Last Notable Activity it's mostly same as Last Activity.







Data Preparation

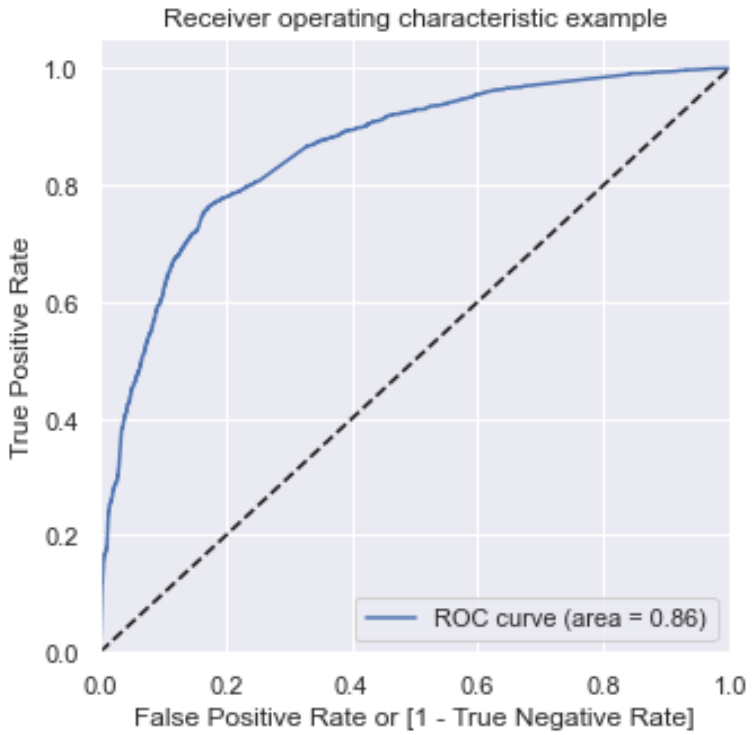
- Created the dummy variables for all the categorical columns.
  - Lead Origin
  - Lead Source
  - Specialization
  - What is your current occupation
  - What matters most to you in choosing a course
- Used standard scalar to scale the data for continuous variables.
  - Total Time Spent on Website
  - Page Views Per Visit
  - Total Visits
- Split the data into train and test data sets with 70% and 30% as proportions, respectively.

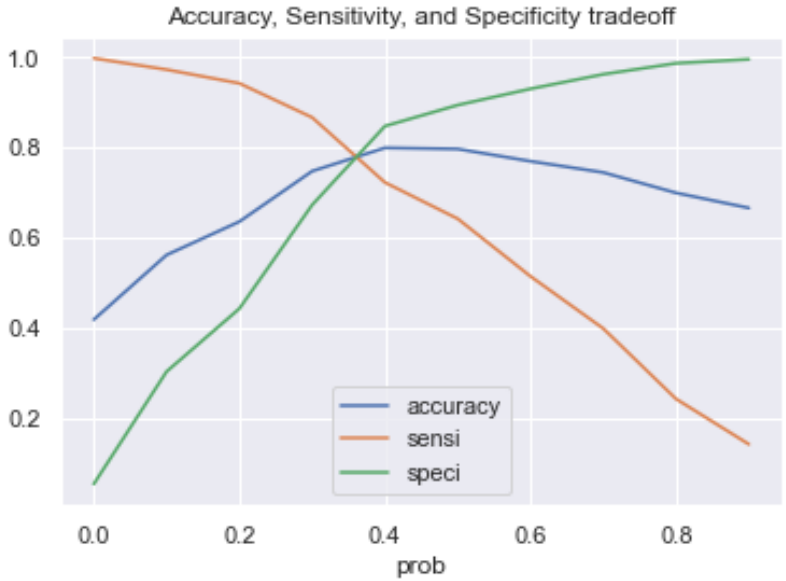
Model Building

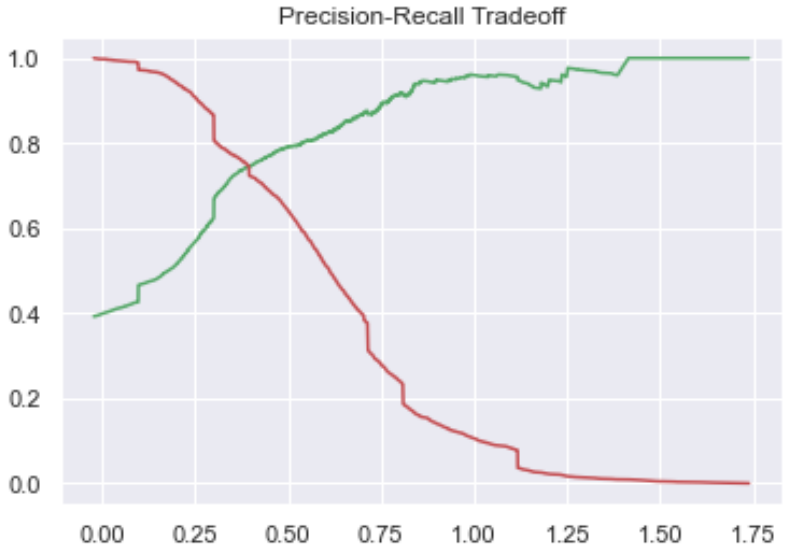
- Using RFE performed feature elimination and selected 20 most impactful variables.
- Analysing VIF scores and p-values eliminated features with VIF score is > 5 or p-value is above 0.05.

Model Evaluation

- In ROC Curve, we can see that 86% of the area is under the curve, which shows that the model is good.
- From the Accuracy, Sensitivity, and Specificity tradeoff, the optimal cut off point is 0.33
- Overall Model Accuracy, Sensitivity, Specificity is ~80%







Conclusion / Final Inferences

- We have assigned a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.
- Company can Target the customers with higher lead score as they have a higher conversion chance.
- Below is the list of characteristics of potential leads, business can focus on these customers to improve their conversion rate.
  1. Leads that are originated through the lead add form.
  2. Customers who are working professionals.
  3. Customers who had spent more time on the course web page.
  4. Also, the leads with Olark Chat as the source of the lead

- Below is the list of Features that are most significant in identifying Hot Leads:
  - Top 3 Features:
    - When the lead origin is Lead add format.
    - When the Occupation is Working Professional
    - The total time spent on the Website.
  - Other Important Features:
    - Total number of visits.
    - When the lead source was Google and Olark Chat
    - When Lead Origin is Landing Page Submission