Lead Score Case Study

BY:

- DAKSHA KANUNGO
- SHUBHAM SINHA

Problem Statement & Business Objective

Problem Statement :

- An education company named X Education sells online courses to industry professionals.
- Company faces the challenge of poor Lead Conversion rate as it gets a lot of leads on website, but very few conversions. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- 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.

Business Objective:

- Company 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.

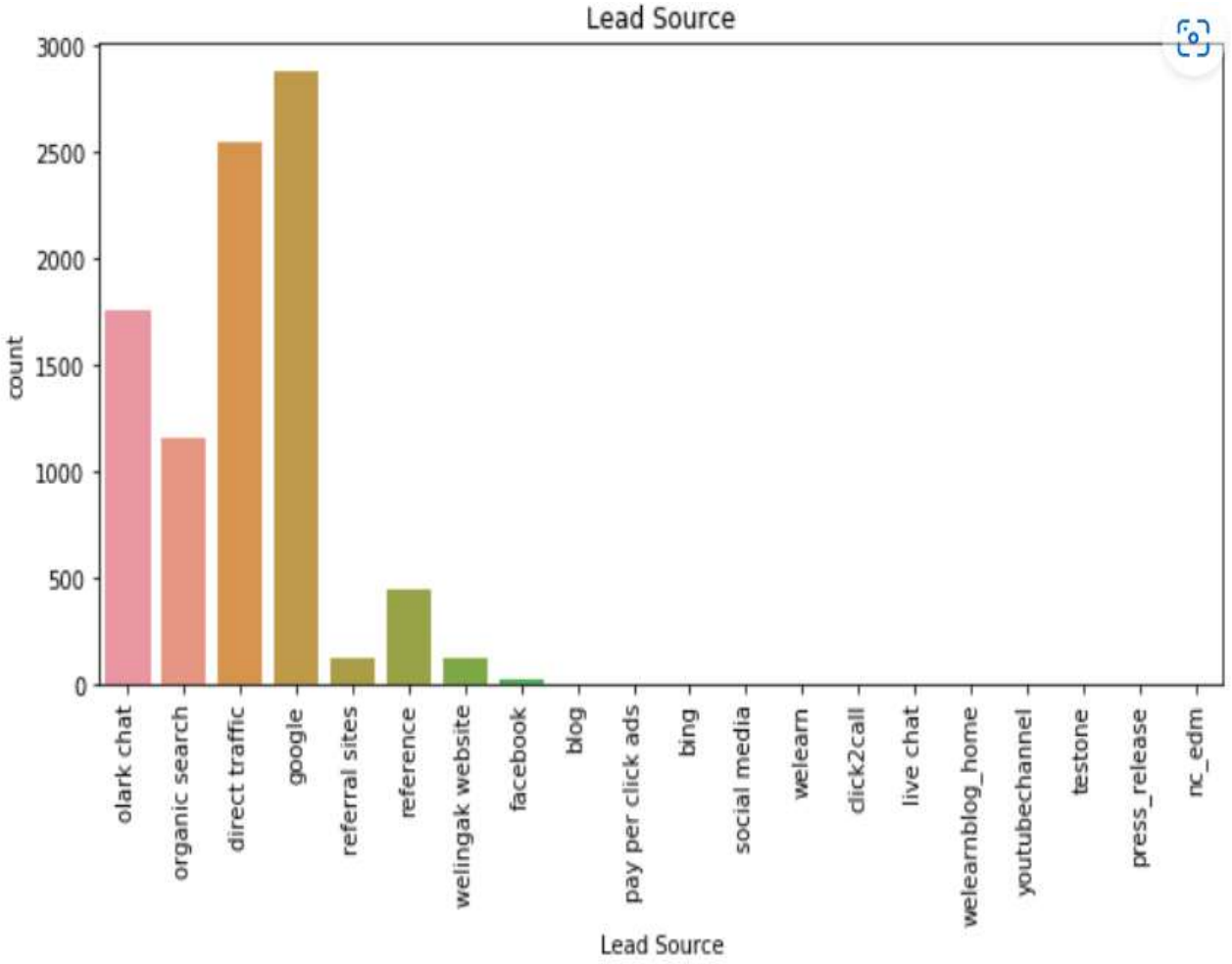
Solution Methodology

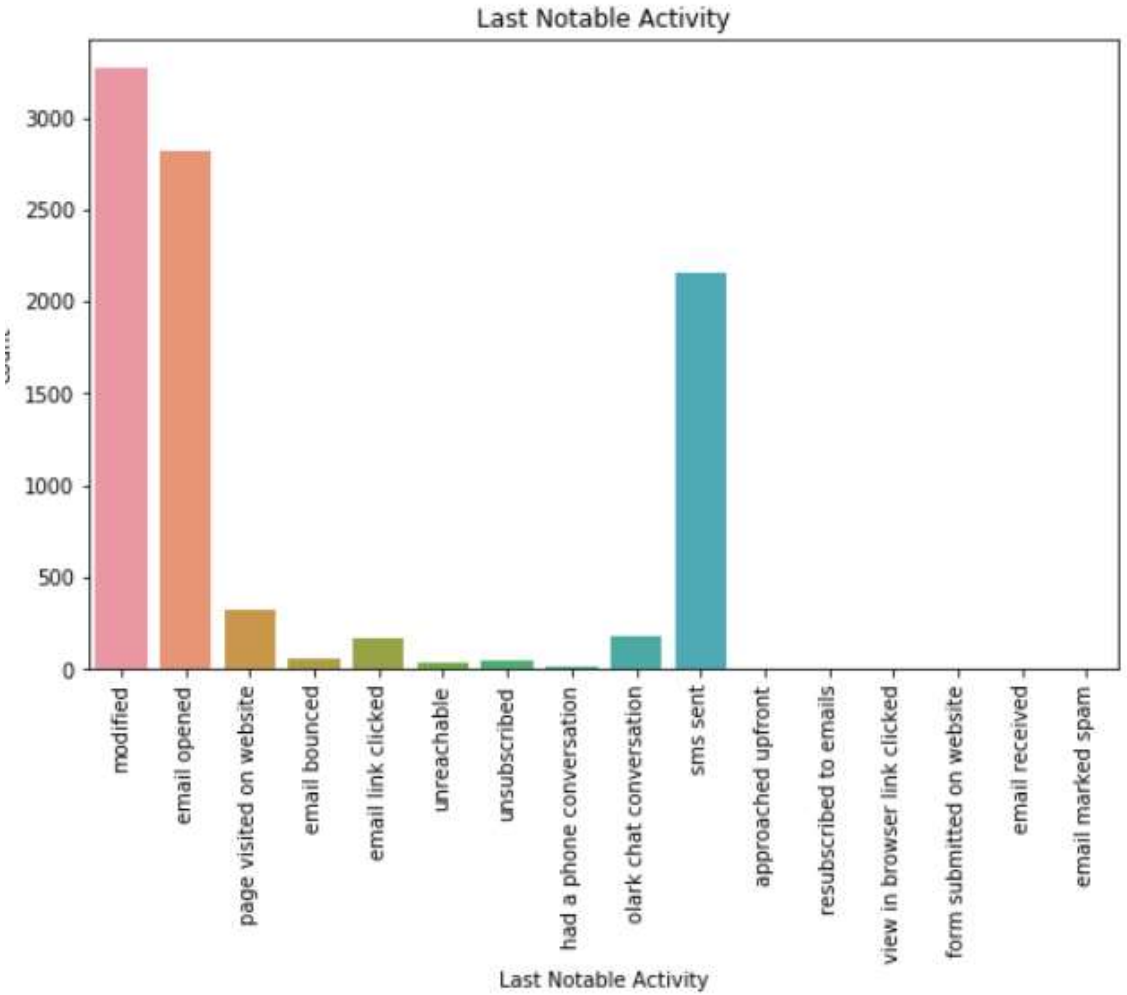
1. Loading Dataset and understanding variables.
2. Data cleaning and data manipulation :
   - Replaced 'Select' with NaN.
   - Handled Missing Values - Dropped Columns with higher percentage of missing values and imputing values where necessary.
   - Checked % of data which could be lost in case of removing rows with NULL values and removed rows with NULL values since percentage was low.
3. Exploratory Data Analysis :
   - Univariate Analysis
   - Bivariate analysis
4. Added Dummy Variables
5. Train Test split
6. Model Building
7. Model Prediction
8. Model Evaluation
9. ROC Curve
10. Prediction on test set

Data Cleaning

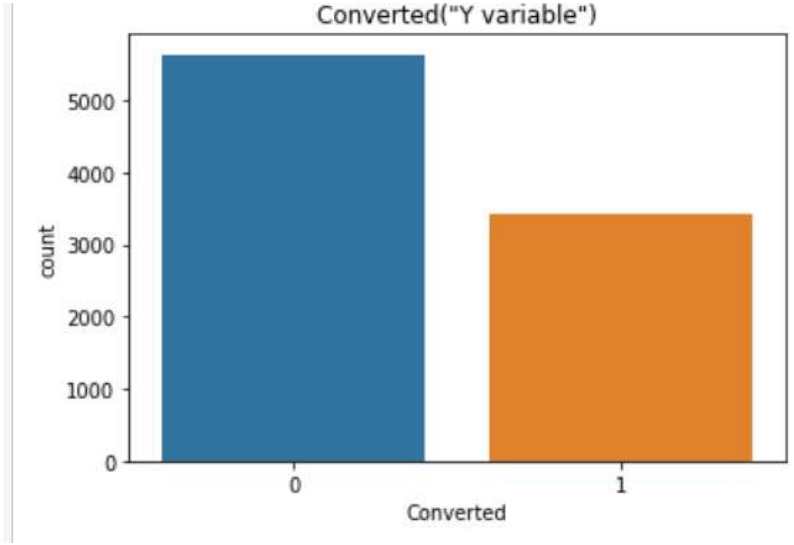
1. Initial data structure :
    1. # Rows = 9240
    2. # Columns = 37
2. Replaced ‘Select’ values with NaN, as it means ‘no option selected’.
3. Deleted single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc.
4. Dropped columns having more than 45% of missing values.
5. Replaced NULL values with NaN, in columns where percentage of NULL values was less than 45%.
6. Checked the percentage of data loss if we remove rows with null values. Since the percentage was 1.48, so removed rows with NULL values.
7. Removed ‘Prospect ID’ column as it was having unique values.

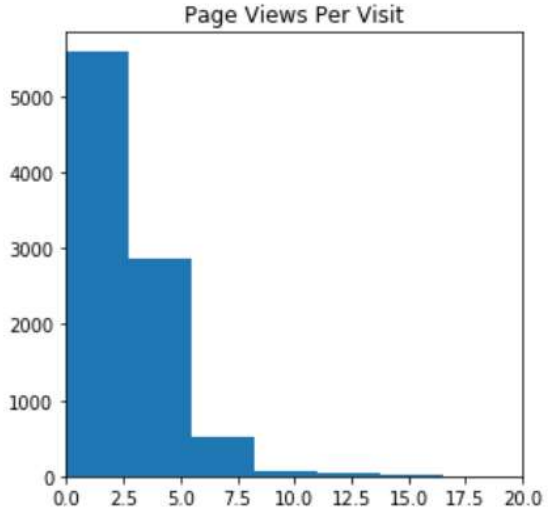
EDA - Univariate Analysis

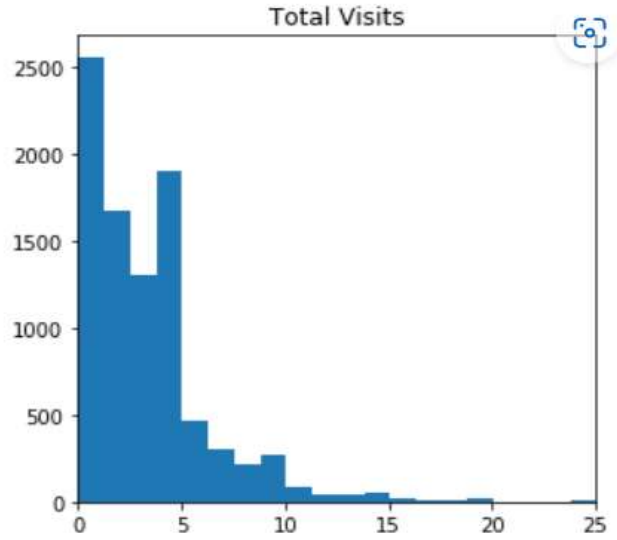


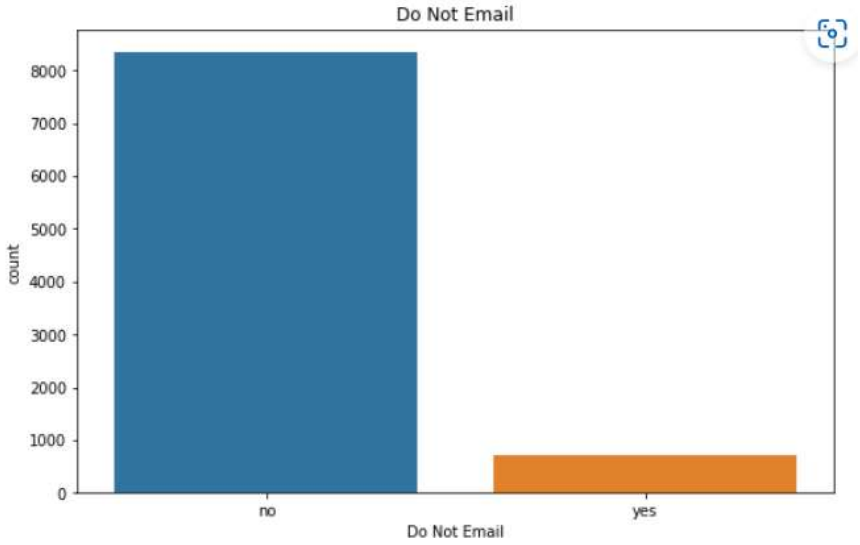


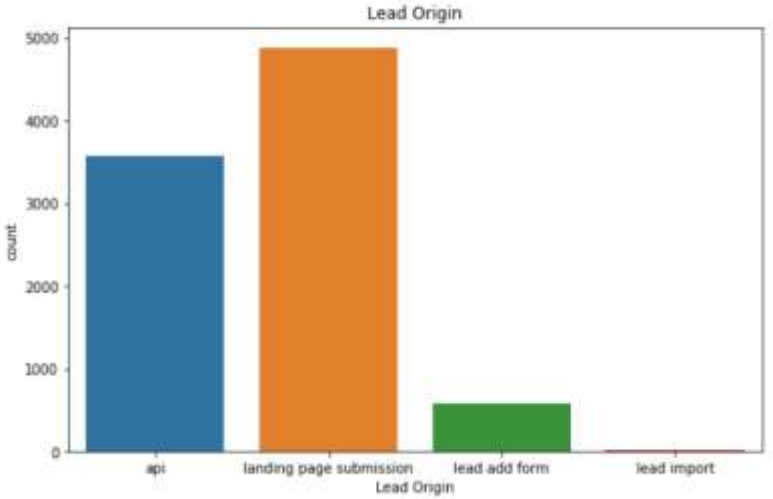
EDA - Univariate Analysis (Cont..)

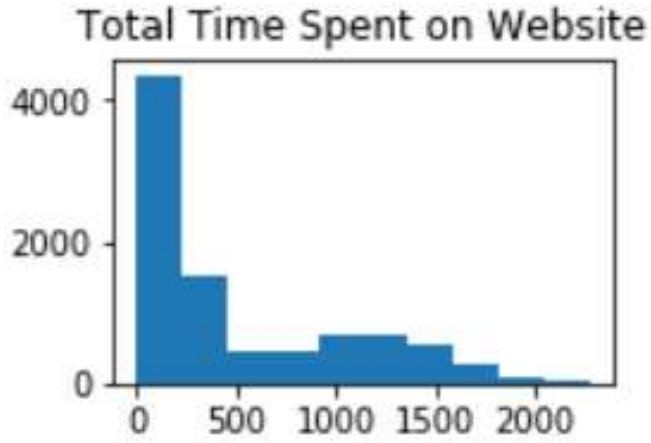




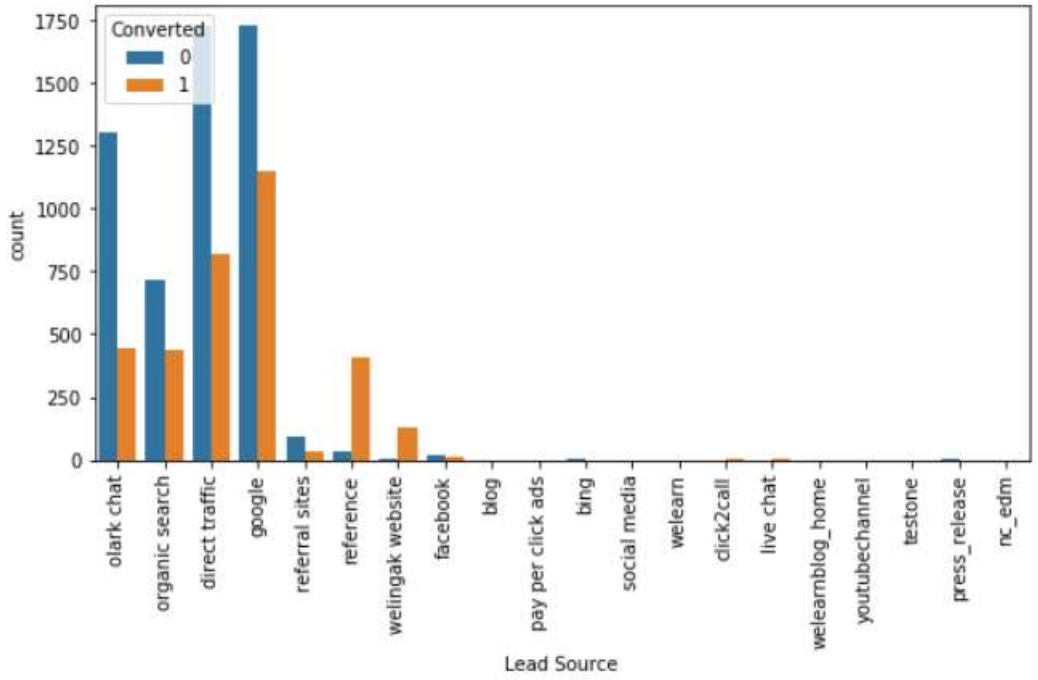


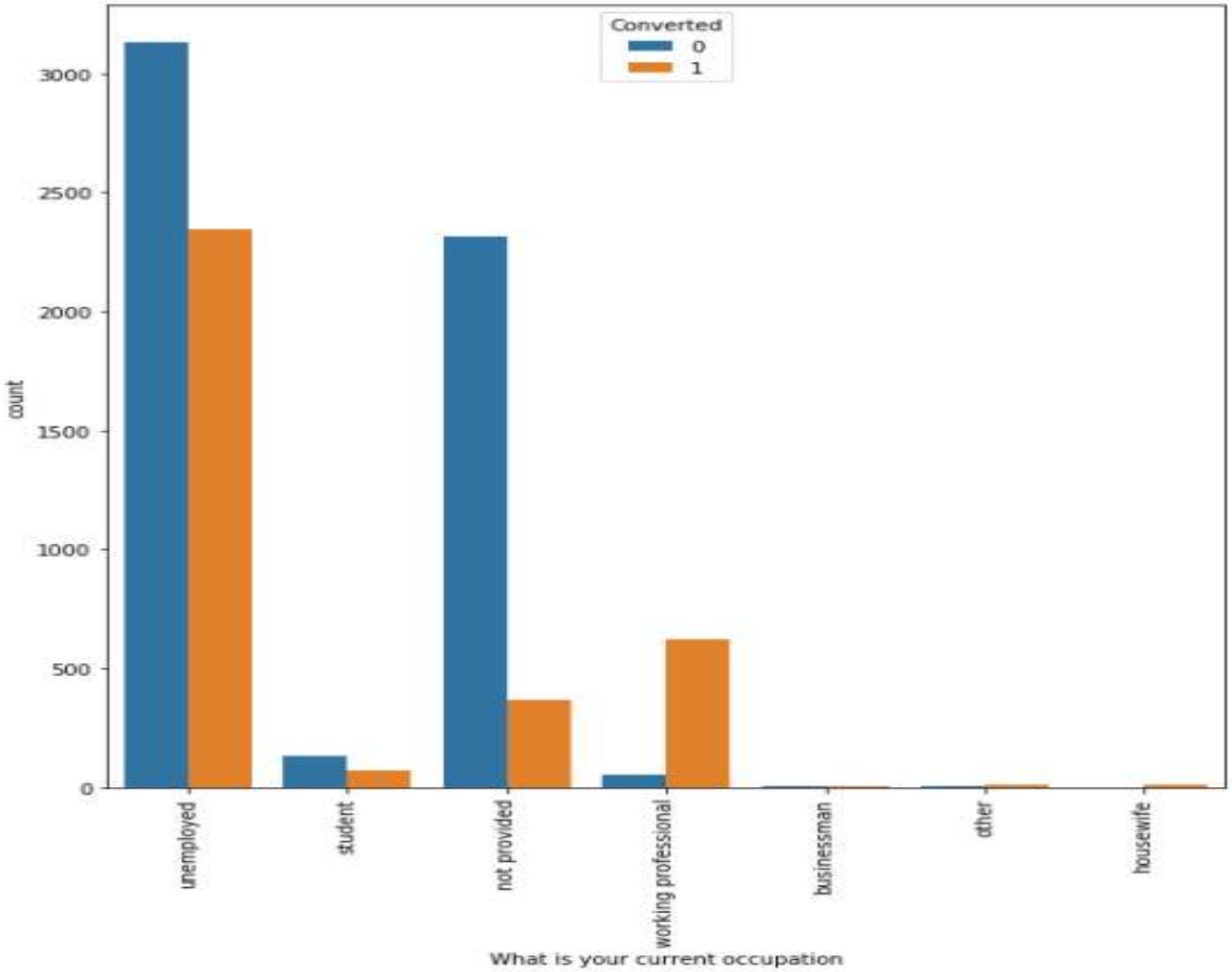




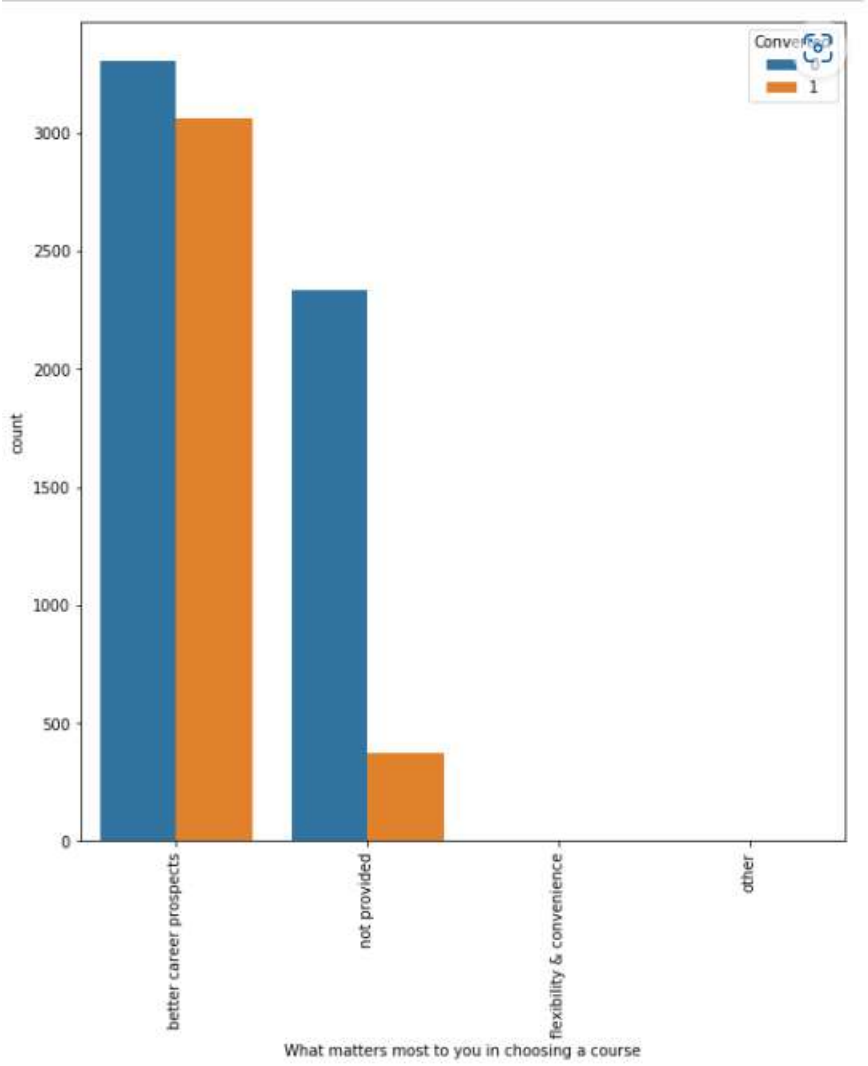


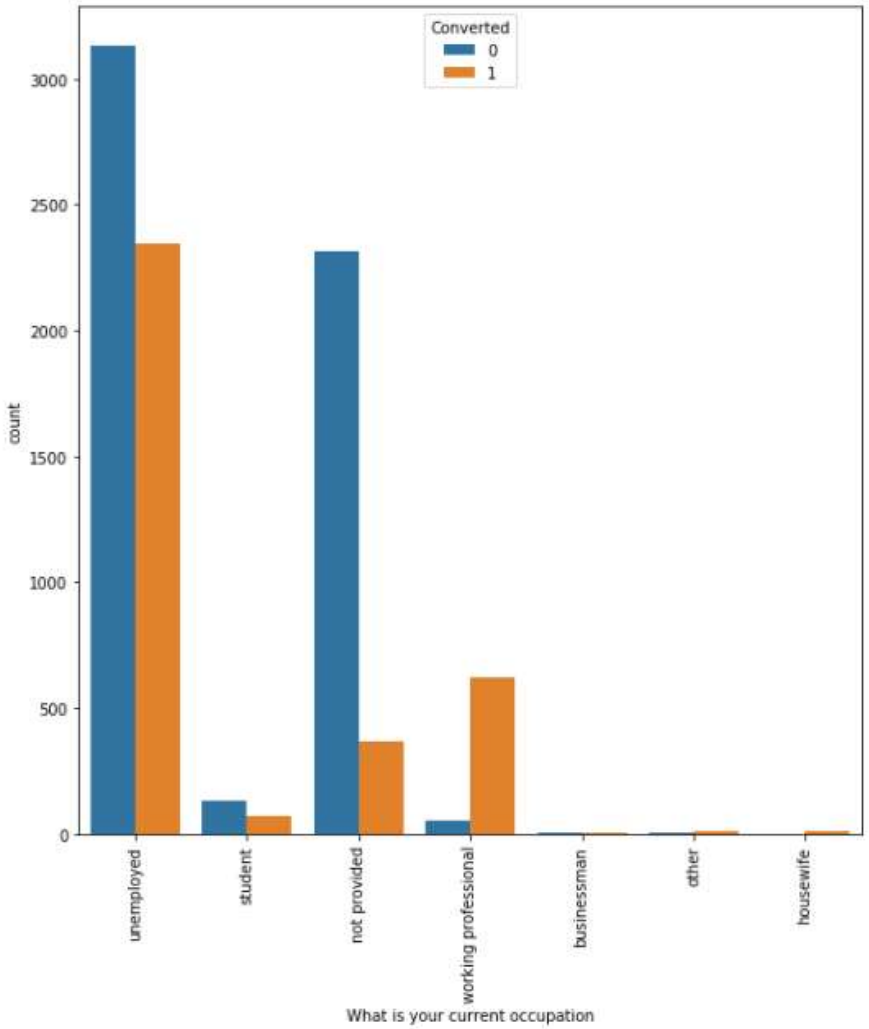
EDA – Relation with Target Variable

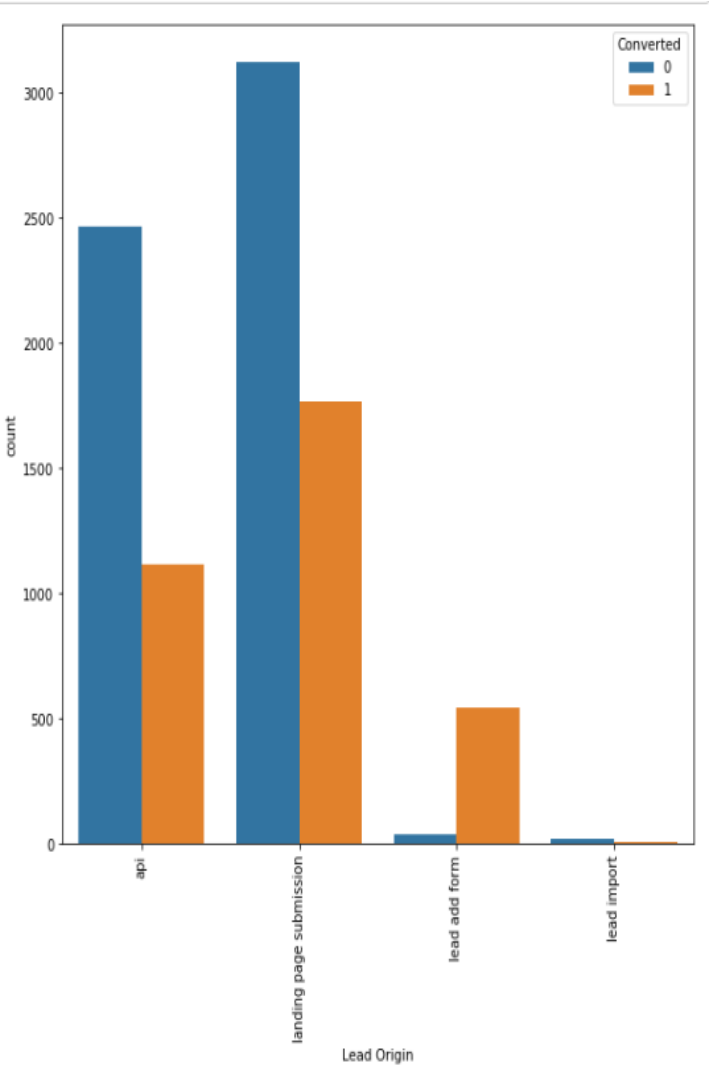




EDA – Relation with Target Variable (Cont..)







Data Conversion, Model Building & Evaluation

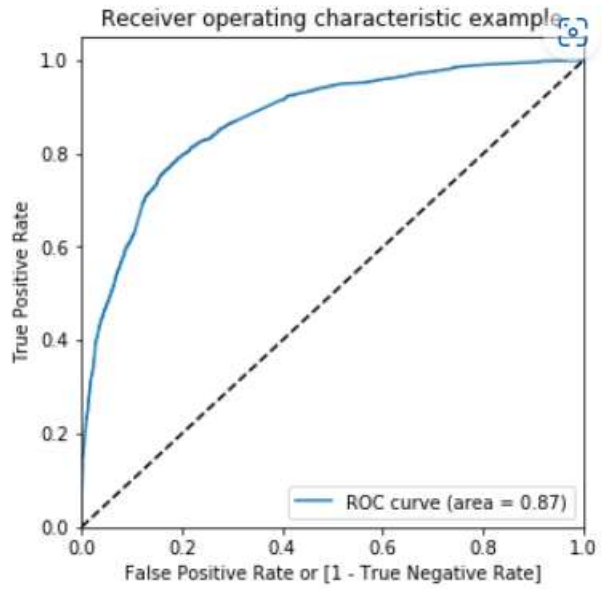
Data Conversion :

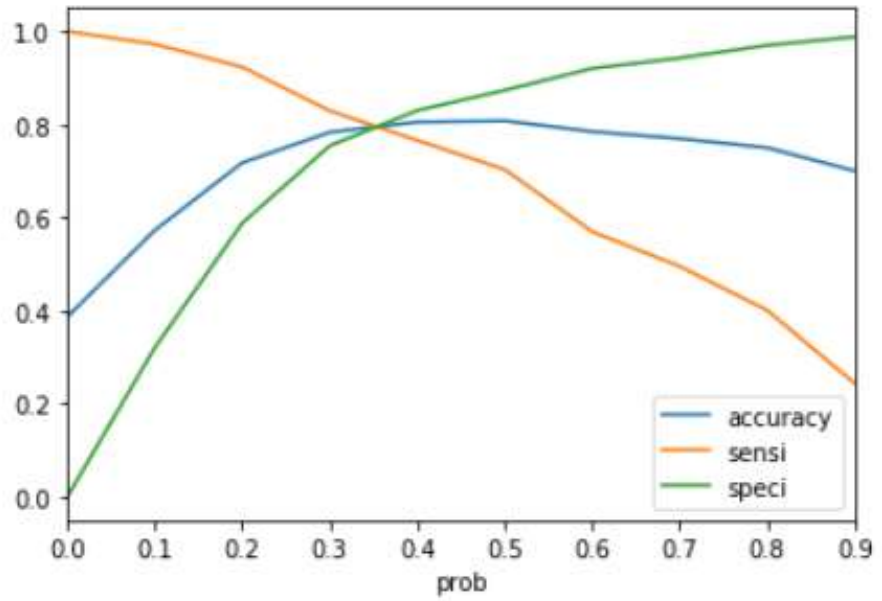
- Normalized numerical variables.
- Created dummy variables for categorical variables.
- Dropped extra columns.

Model Building & Evaluation:

- Divided data into training and test sets using 70:30 ratio.
- Scaled Numerical variables.
- Used RFE for Feature Selection and ran RFE with 15 variables as output.
- Built 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.
- Model Evaluation by checking confusion metrics - value of sensitivity was ~70% and specificity is ~87% with a cutoff of 0.5.
- Overall accuracy of model is ~81%.

ROC Curve





- Used ROC function to get the Optimal cut off point (point where sensitivity and specificity are balanced).
- In first graph, the area under curve is 0.87, which is pretty good value.
- From the second graph it is visible that the optimal cut off is at 0.35.

Conclusion

Looking at the model, it can be concluded that the variables that matter the most in identifying potential leads are:

- 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 SMS sent and 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 grow there business as they have a very high chance to get most of the potential buyers to change their mind and buy their courses.