LEAD SCORING CASE STUDY

Group Members:
1. Prerna Gupta
2. Siddharth Aher
3. Ruchi Rawat

PROBLEM STATEMENT

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. 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. The typical lead conversion rate at X education is around 30%. Now, although X Education gets a lot of leads, its lead conversion rate 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.

BUSINESS GOAL

Build a Logistic Regression model to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

SOLUTION METHODOLOGY

- Data Cleaning & Data Manipulation
  - Check and Handle Duplicate Data.
  - Check and Handle missing values and NA values.
  - Drop columns, if it contains large amount of missing values and not useful for analysis.
  - Imputation of the values if necessary.
  - Check and handle outliers in data
- EDA
  - Univariate Analysis: Value Count, Distribution of Variable, etc.
  - Bivariate Analysis: Correlation Coefficients and Pattern Between Variables, etc.
- Feature Scaling & Dummy Variables and Encoding of Data.
- Classification Technique: Logistic Regression used for the Model Making and Prediction.
- Validation of the Model.
- Model Presentation
- Conclusion and Recommendations

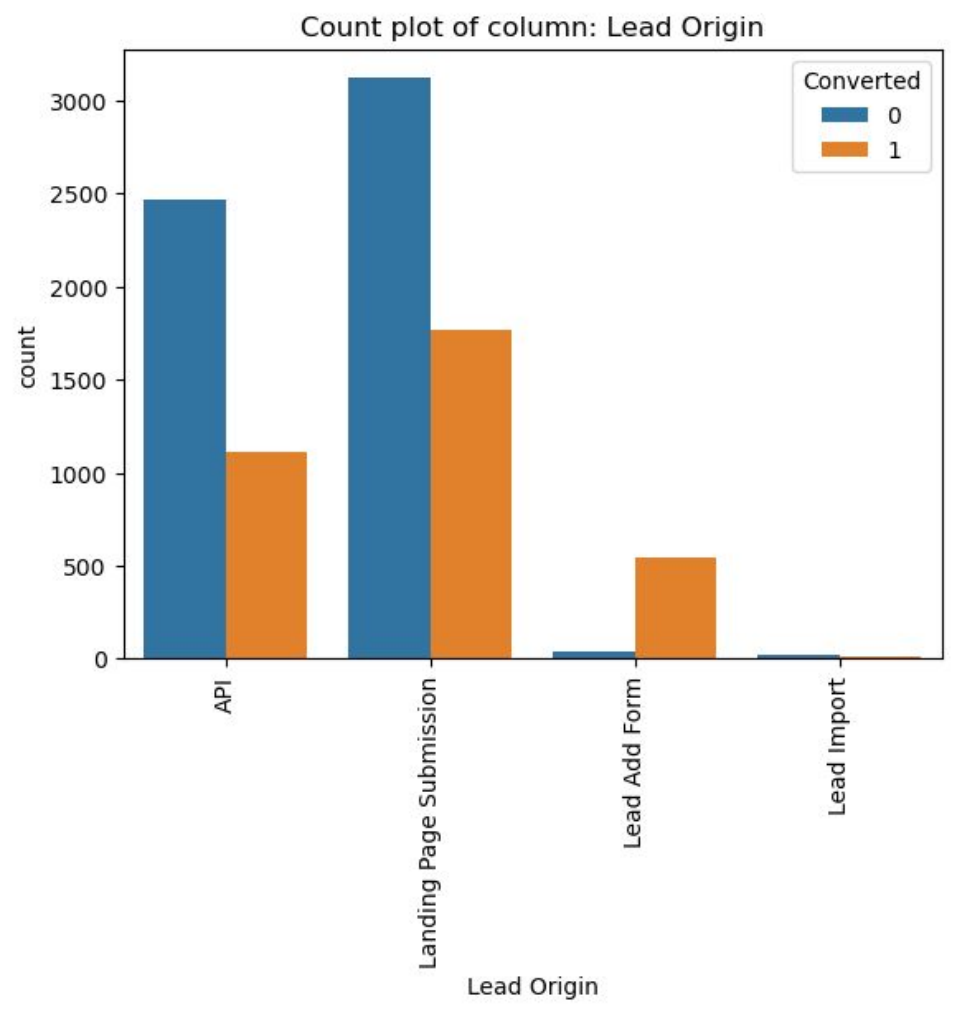
DATA MANIPULATION

- Total number of columns: 37, Total number of rows: 9240
- Many values in some columns were marked as “Select”. There were no such specific option given so we replaced the select values as “Null”.
- Dropped the columns with more than 40% missing values such as “Lead Profile”, “Lead Quality”, “Asymmetric profile score”, etc.
- The following columns’ null values were replaced by most appeared values in their respective columns. The columns are: “City”, “Tags”, “What matters most to you in choosing a course”, “What is your current occupation”, “Country”.
- In column “Specialization”, it may be possible that the leads may leave this column blank if he/she is a student or their specialization is not mentioned in options given.
- Rest of the columns had less then 2% null values so they were removed.

EXPLORATORY DATA ANALYSIS

Univariate and Bivariate Analysis

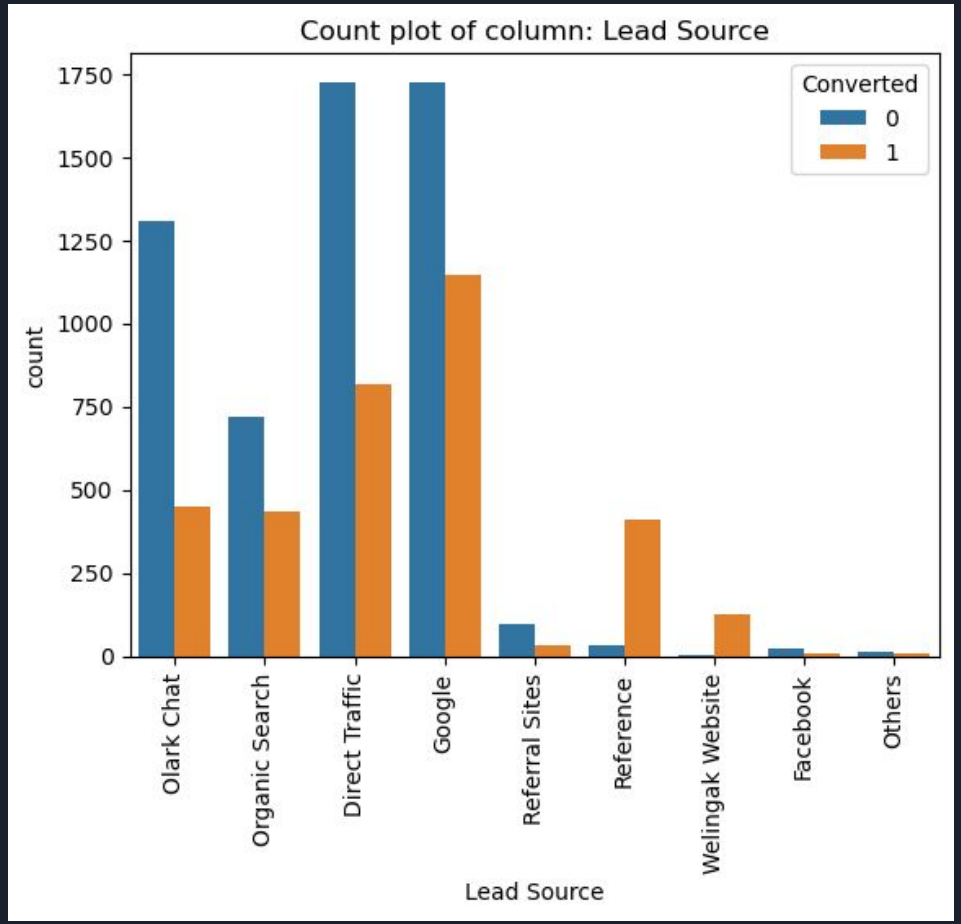
- Converted Value of 1 means that the lead was converted and 0 means it was not.
- We have 38 % leads data that was converted and 62 % that was not.



Lead Origin

1. API and Landing Page Submission have a bit lower conversion rate but count of lead originated from them are considerable.

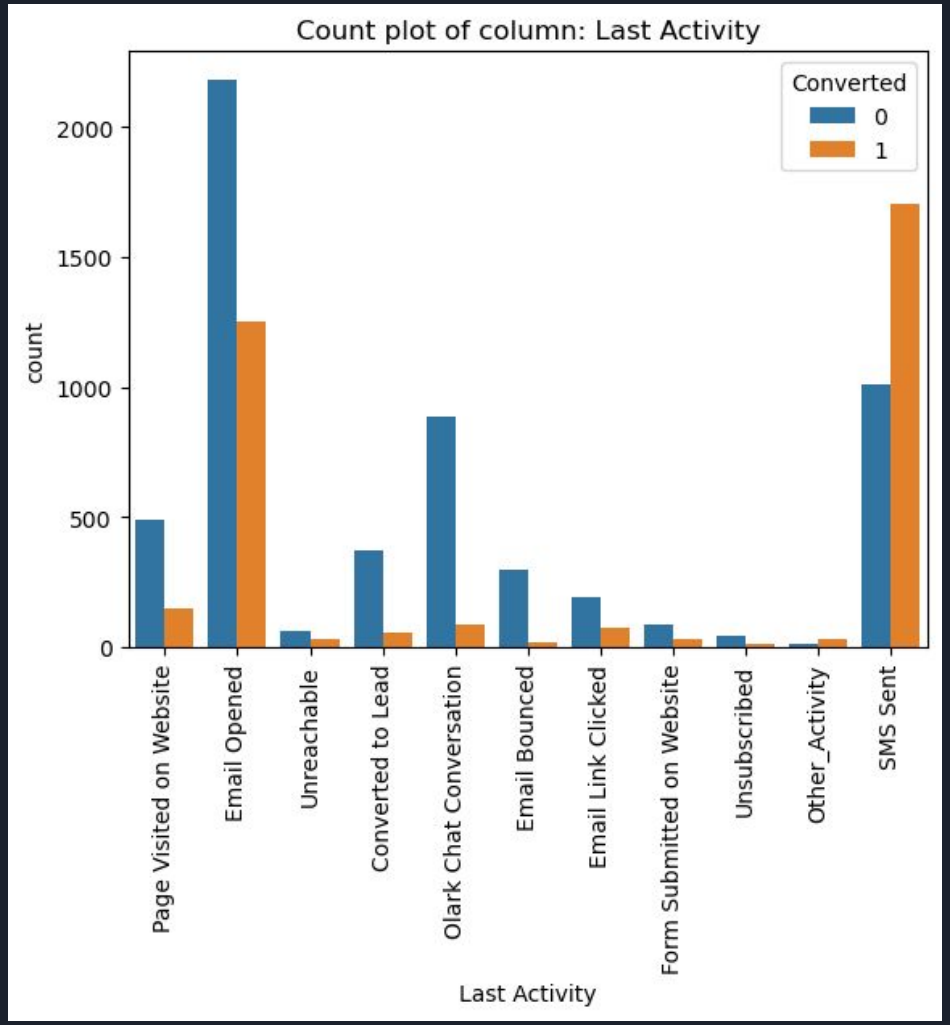
2. Lead Add Form has more than 90% conversion rate but count of leads are not very high.



Lead Source

1. The conversion of leads from Welingak website is maximum.

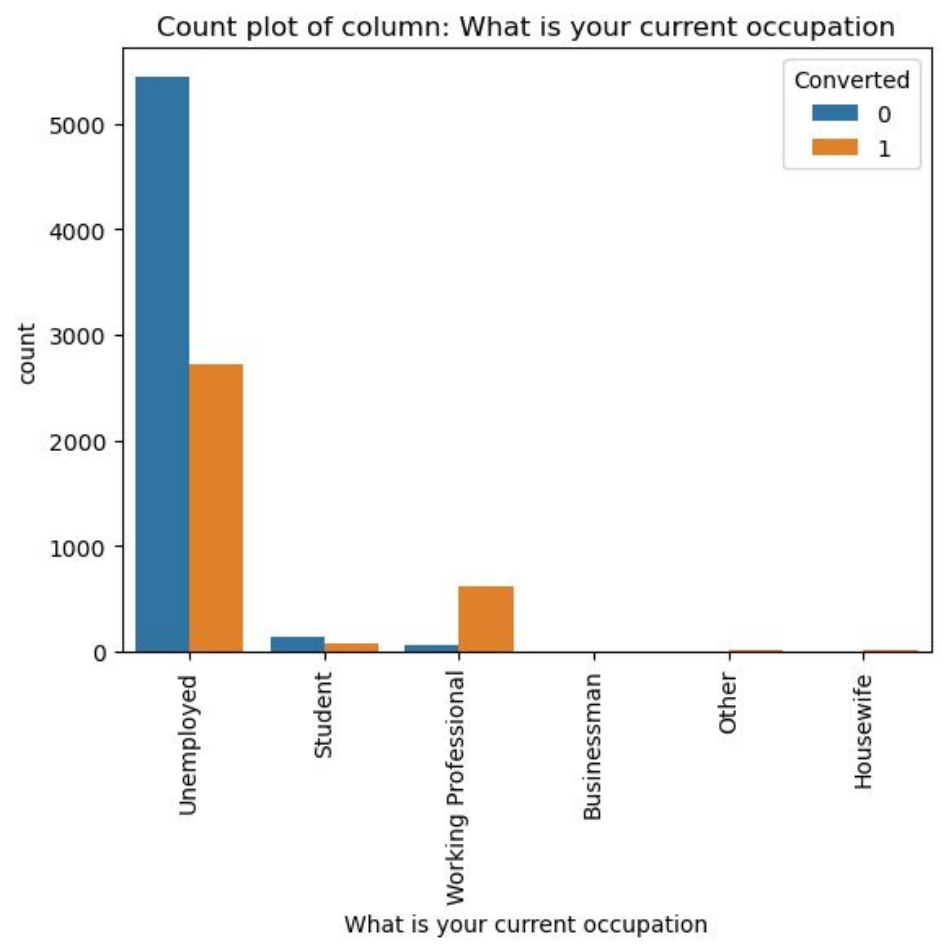
2. The leads generated from google and direct link is high but conversion is minimum.



Last Activity

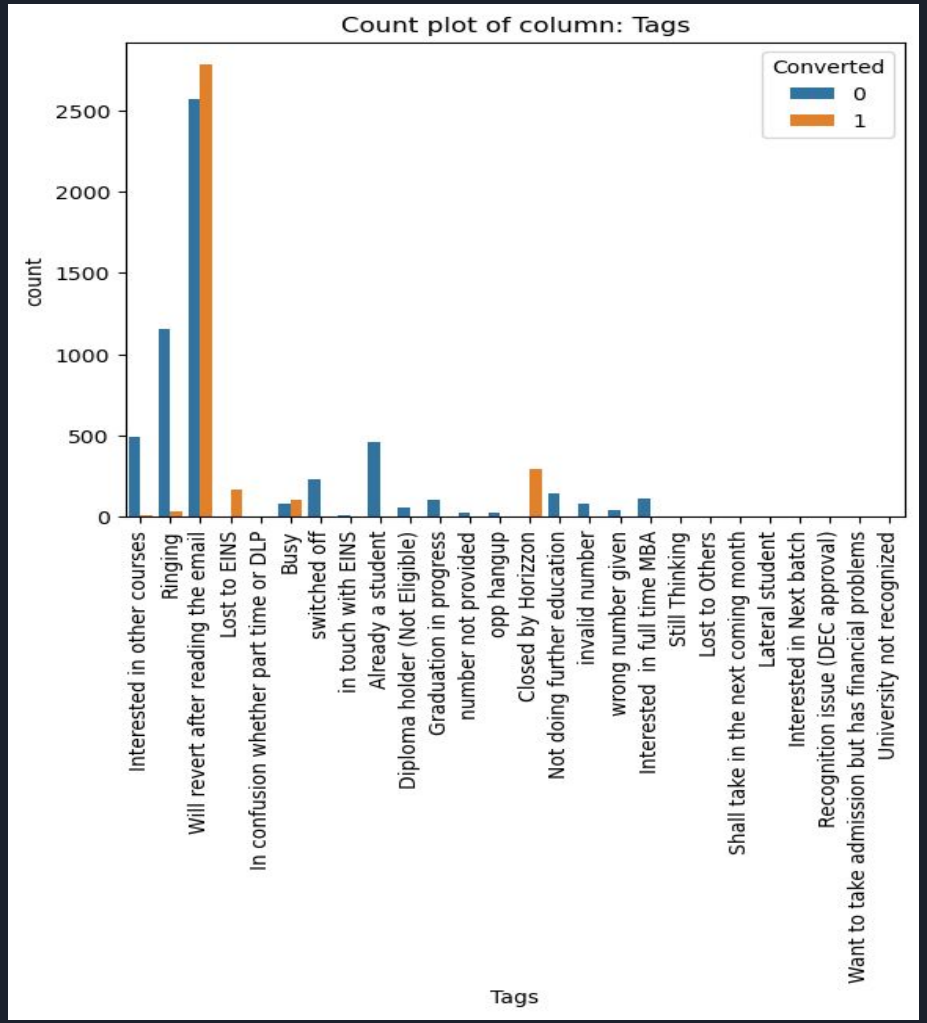
1. Conversion rate is maximum for SMS sent column.

2. Most of the leads have 'Email Opened' as the last activity.



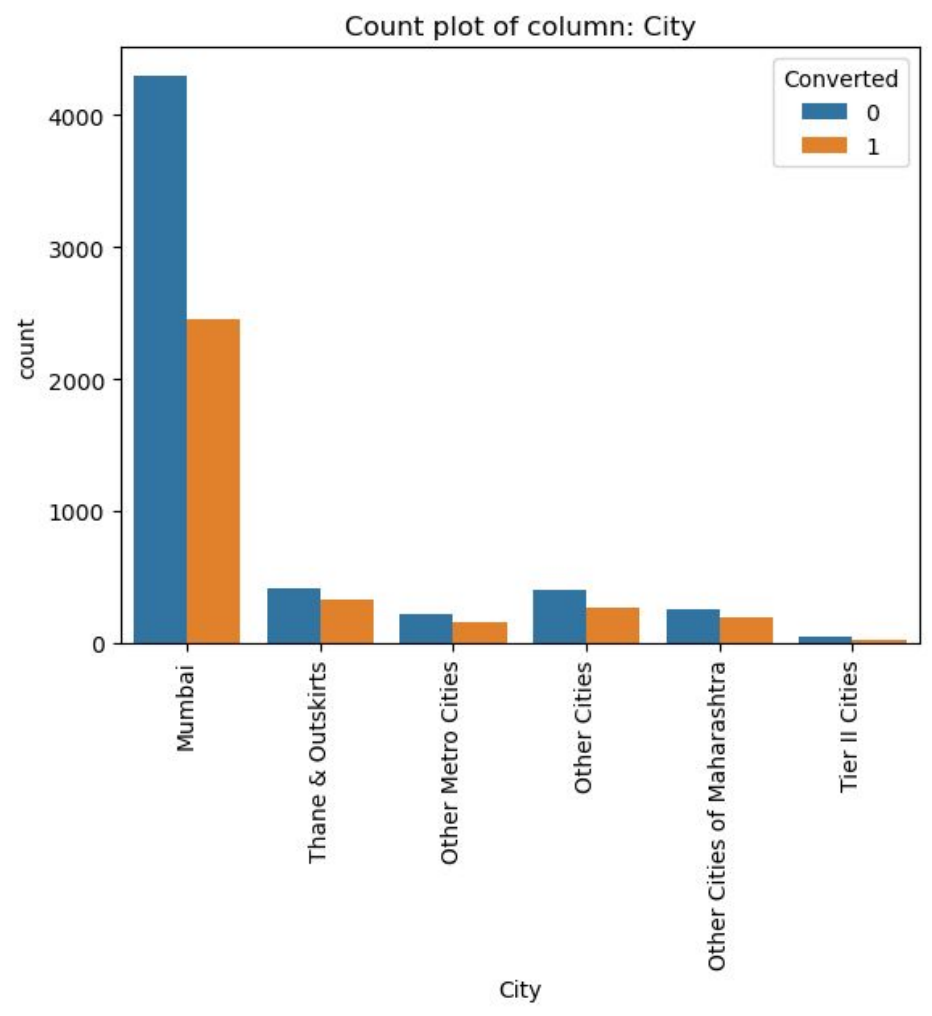
Current Occupation

- We can see that working professional has higher conversion rate as compared to unemployed which has maximum number of leads.



Tags

- We can see that “will revert after reading the Email” have maximum conversion rate



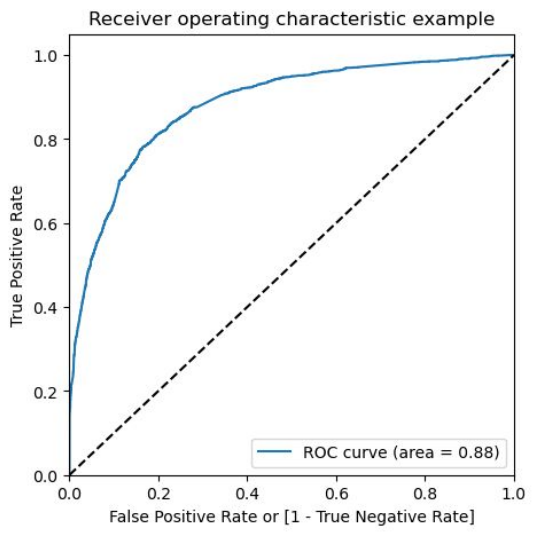
City

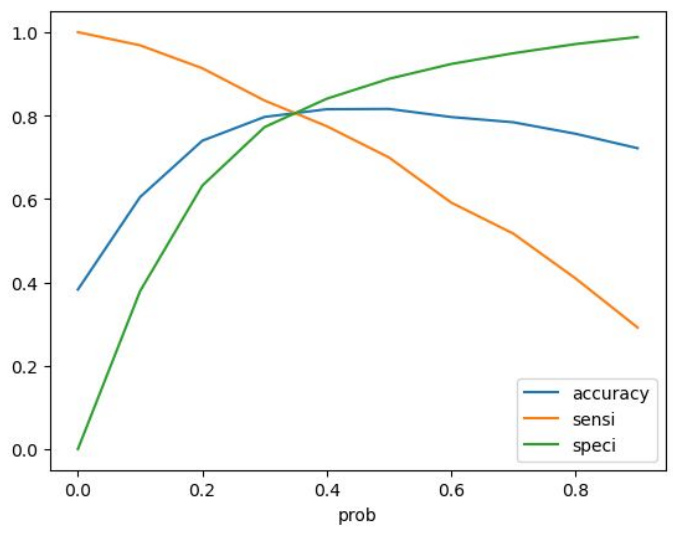
- Mumbai city has most converted leads.

MODEL BUILDING

- Splitting the Data into Training and Testing Sets
- The First basic test for regression is performing a train test split, we have chosen 70:30 ratio.
- Use RFE for feature selection.
- Running RFE with 20 variables as output.
- Building Model by removing the variable whose p-values is greater than 0.05 and VIF value is greater than 5.
- Predictions of Test Data set.
- Overall Accuracy 80%

ROC CURVE





- Finding Optimal Cutoff Point
- Optimal Cutoff probability is that
- Probability where we get balanced sensitivity and specificity
- From second graph it is visible that the optimal cut off is at 0.35

CONCLUSION

1. More budget/spend can be done on Welingak Website in terms of advertising, etc.
2. Incentives/discounts for providing reference that convert to lead, encourage to provide more references.
3. Working professionals to be aggressively targeted as they have high conversion rate and will have better financial situation to pay higher fees too.