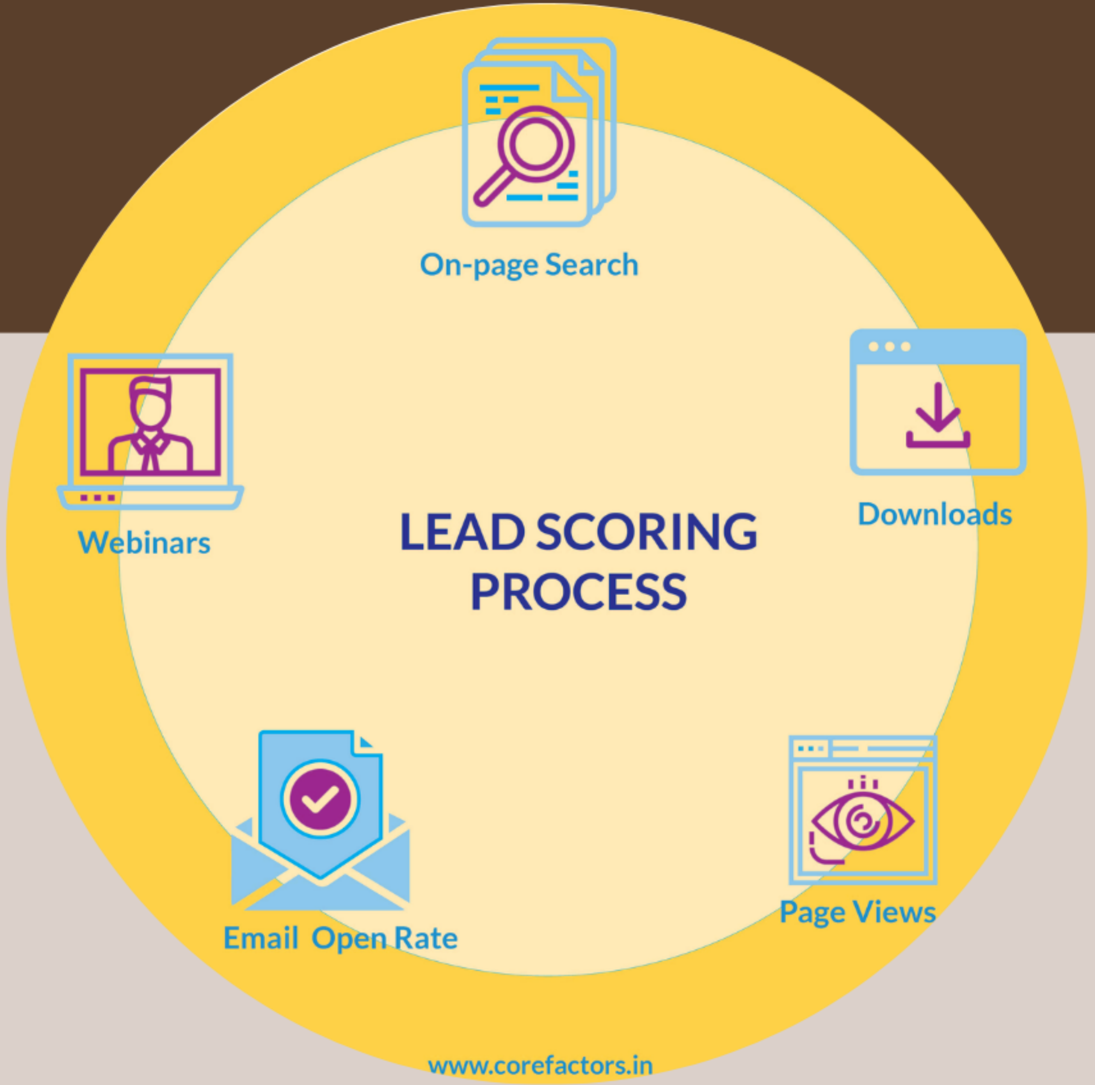
LEAD SCORING CASE STUDY

By

Satyam Khorgade



PROBLEM STATEMENT

▶ X Educationsells online courses to industry professionals.

▶ X Education gets a lot of leads, its lead conversion rate is very poor. 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 knownas ‘Hot Leads’.

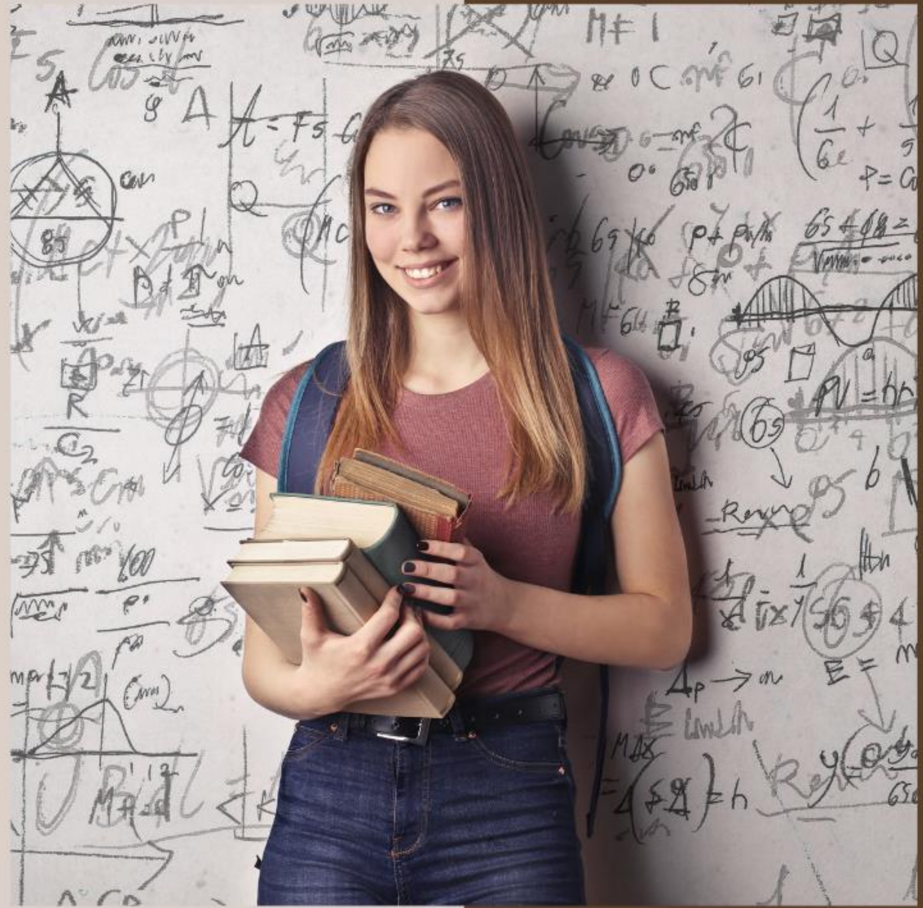
▶ If they successfully identify this set of leads, the lead conversion rate should go upas the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

BUSINESS OBJECTIVE

▶ X education 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

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 a large number of missing values and are not useful for the analysis.

4. Imputation of the values, if necessary.

5. Check and handle outliers in data.

Exploratory Data Analysis (EDA)

1. Univariate data analysis: value count, distribution of variables, etc.

2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.

3. Feature Scaling & Dummy variables and encoding of the data.

4. Classification technique: logistic regression is used for model making and prediction.

5. Validation of the model.

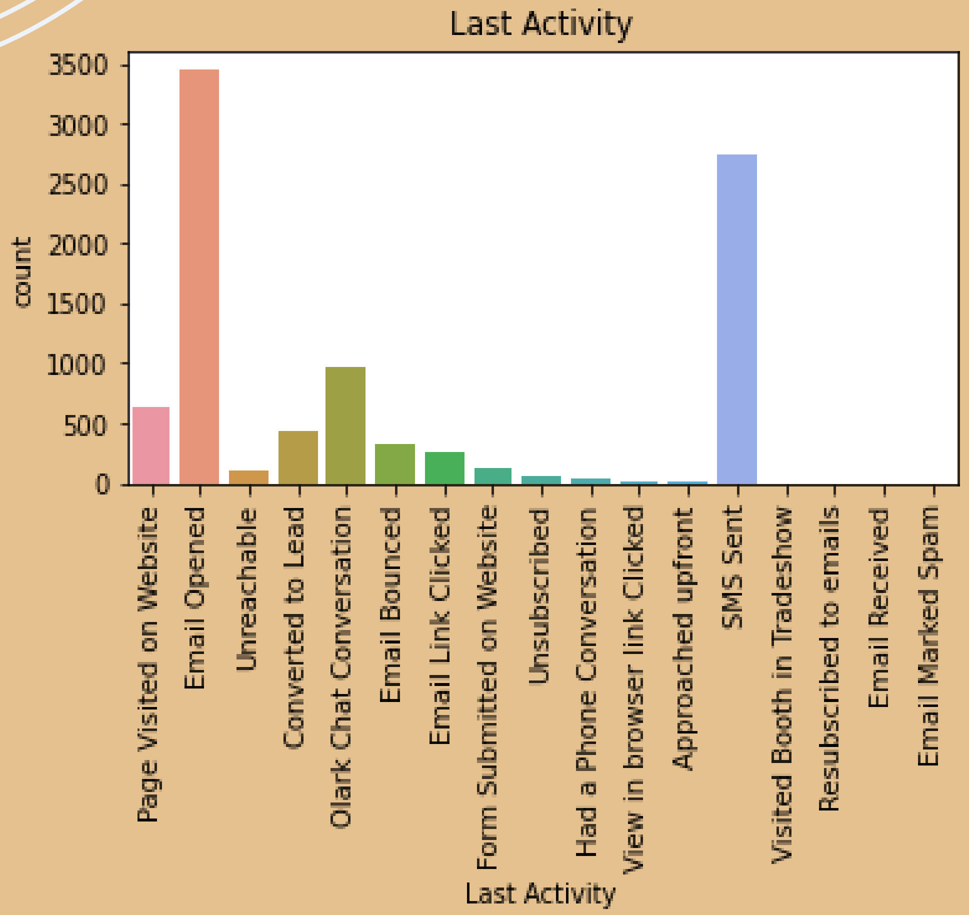
6. Model presentation.

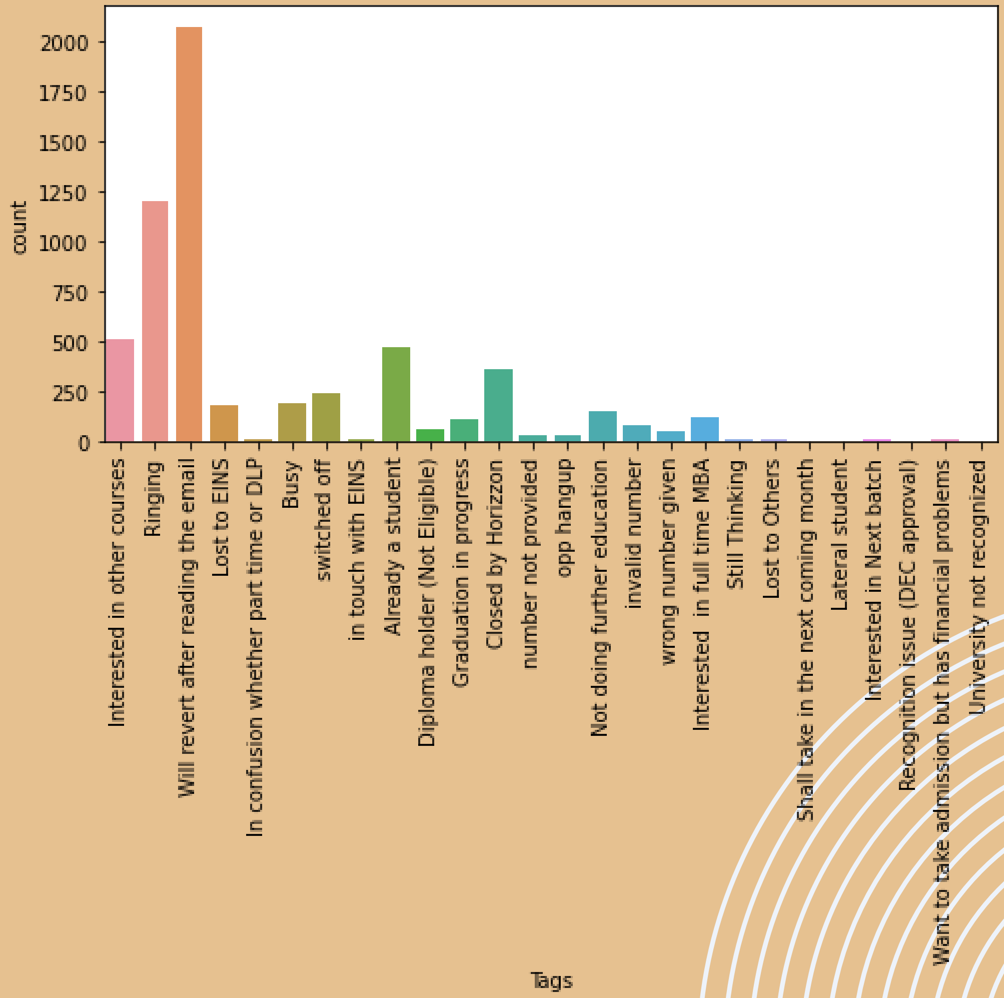
7. Conclusions and recommendations.

DATA MANIPULATION

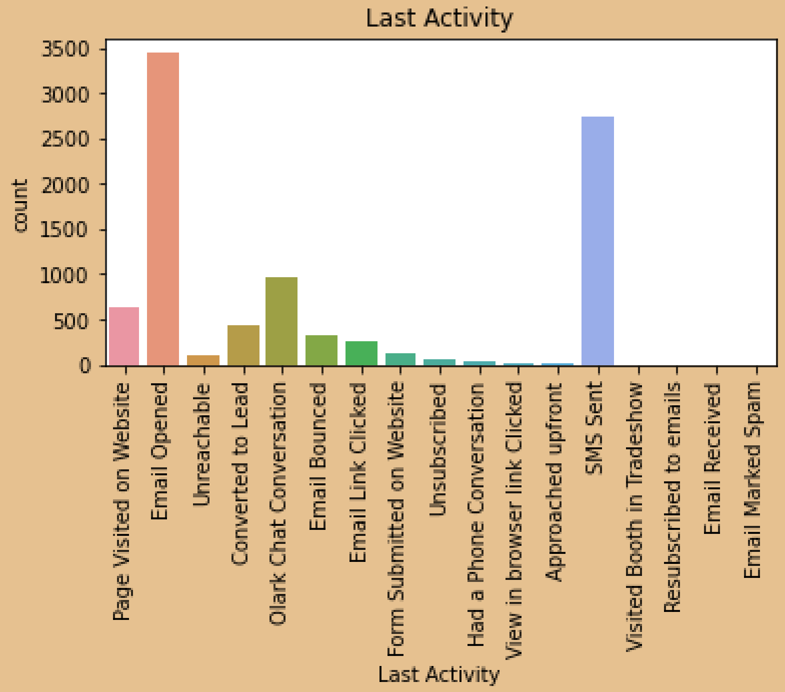
- Total Number of Rows =37, Total Number of Columns =9240.
- Single value features like "Magazine", "Receive More Updates About Our Courses", "Update me on Supply" Chain Content", "Get updates on DM Content", "I agree to pay the amount through cheque" etc. have been dropped.
- Removing the "Prospect ID" and "Lead Number" which is not necessary for the analysis.
- After checking for the value counts for some of the object type variables, we find some of the features which has not enough variance, which we have dropped, the features are: "Do Not Call", "What matters most to you in choosing course", "Search", "Newspaper Article", "X Education Forums", "Newspaper", "Digital Advertisement" etc.
- Dropping the columns having more than 35% as missing value such as 'How did you hear about X Education' and 'Lead Profile'.

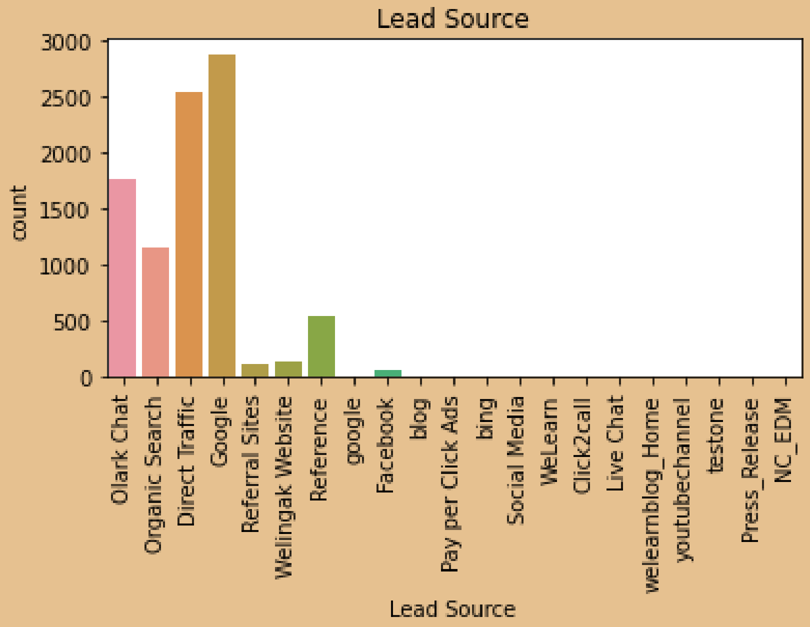
EDA (Exploratory Data Analysis)

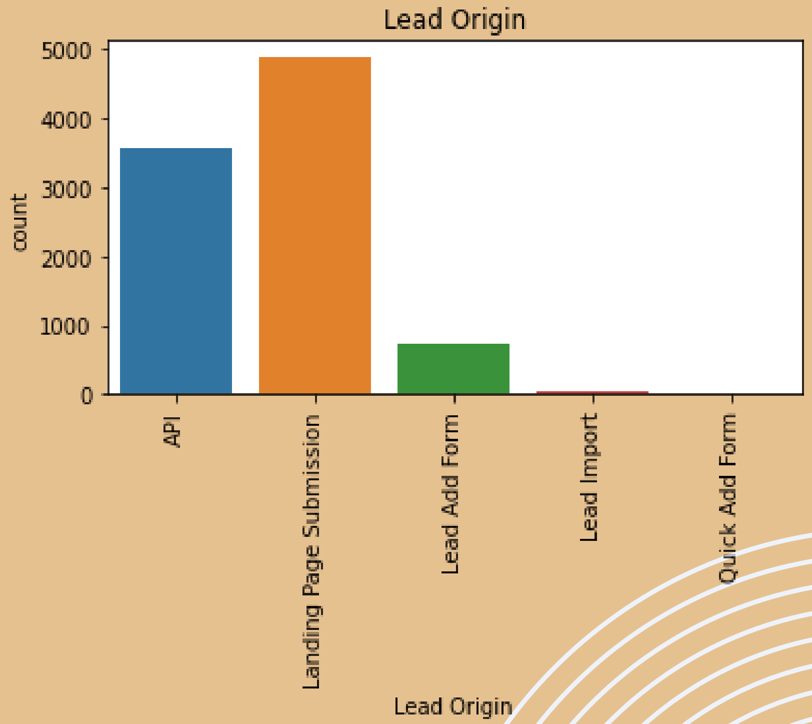




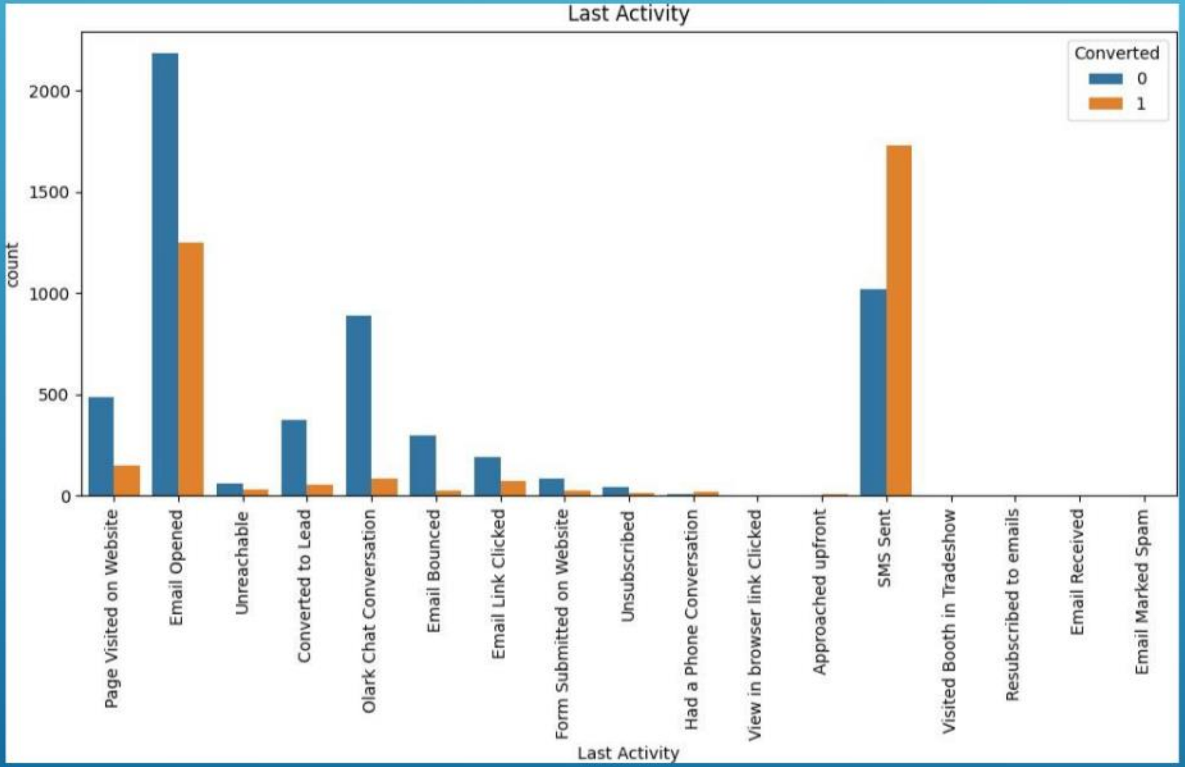
Notable Eda Conclusions

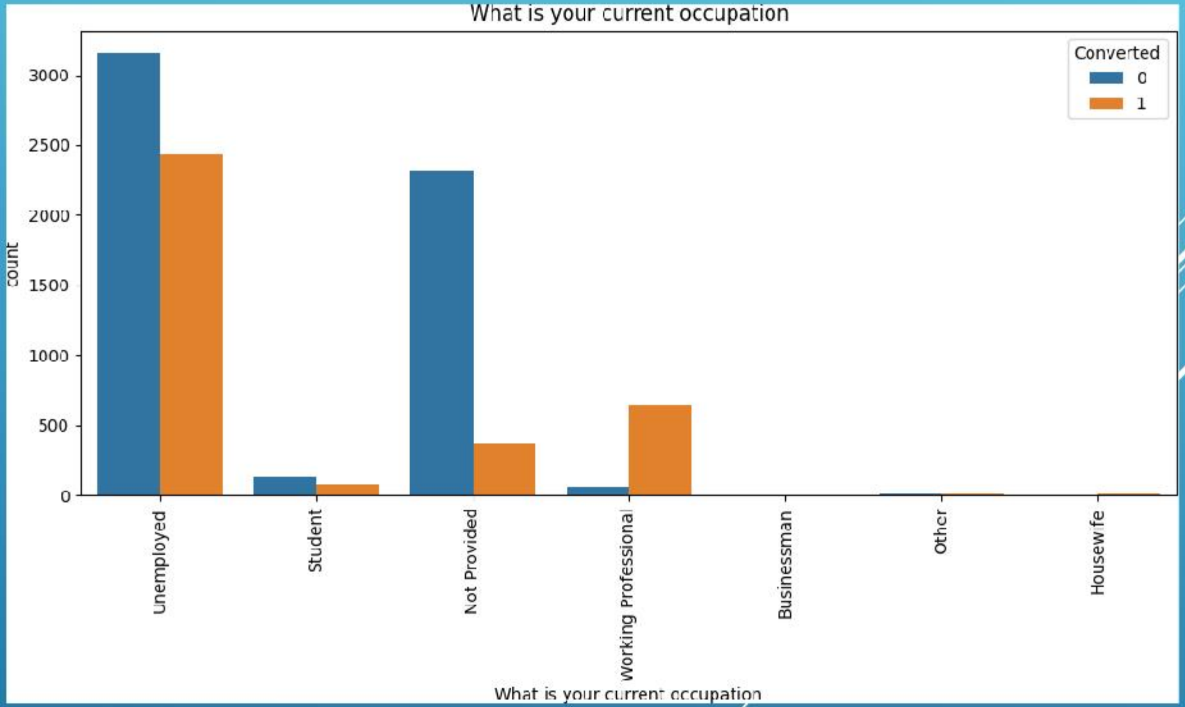






NOTABLE EDA CONCLUSIONS (CONT.)

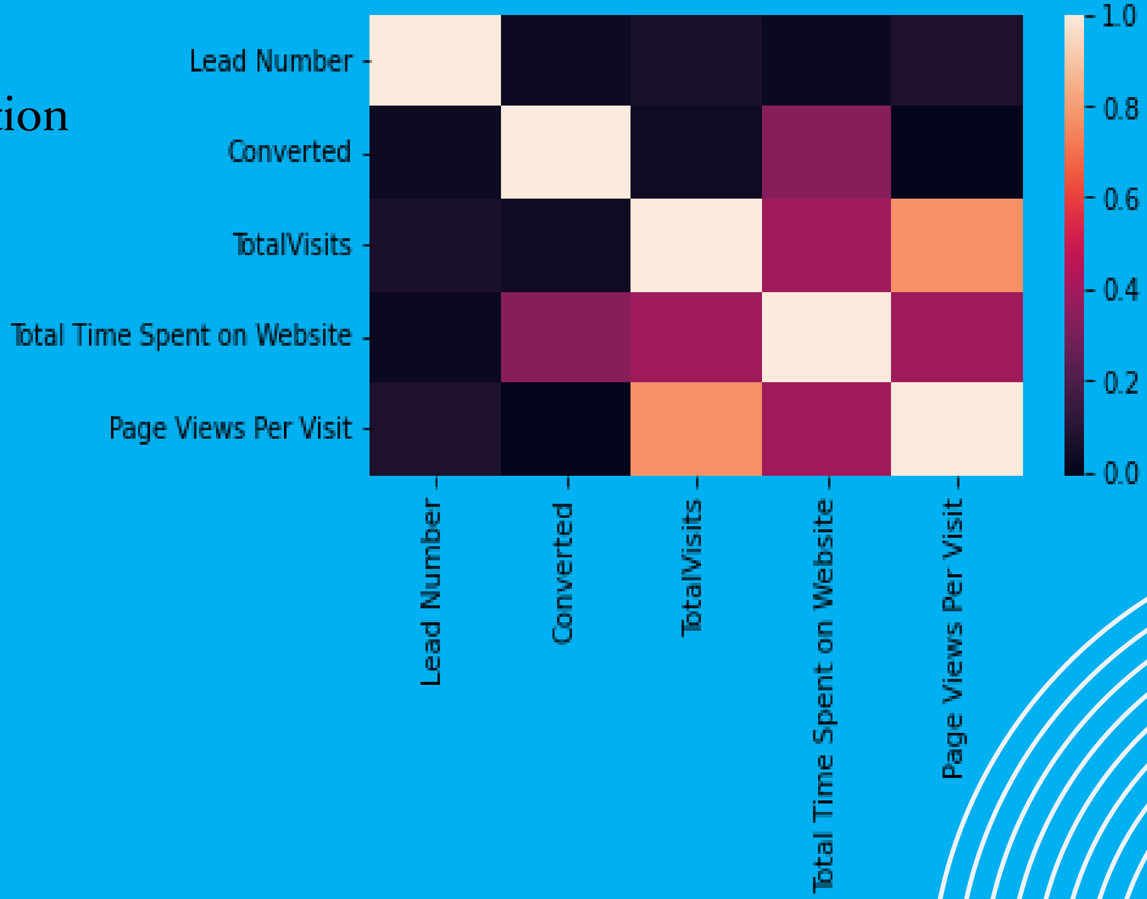




CORELATION MATRIX

This correlation matrix displays high correlation between 'Converted' and 'Lead number'

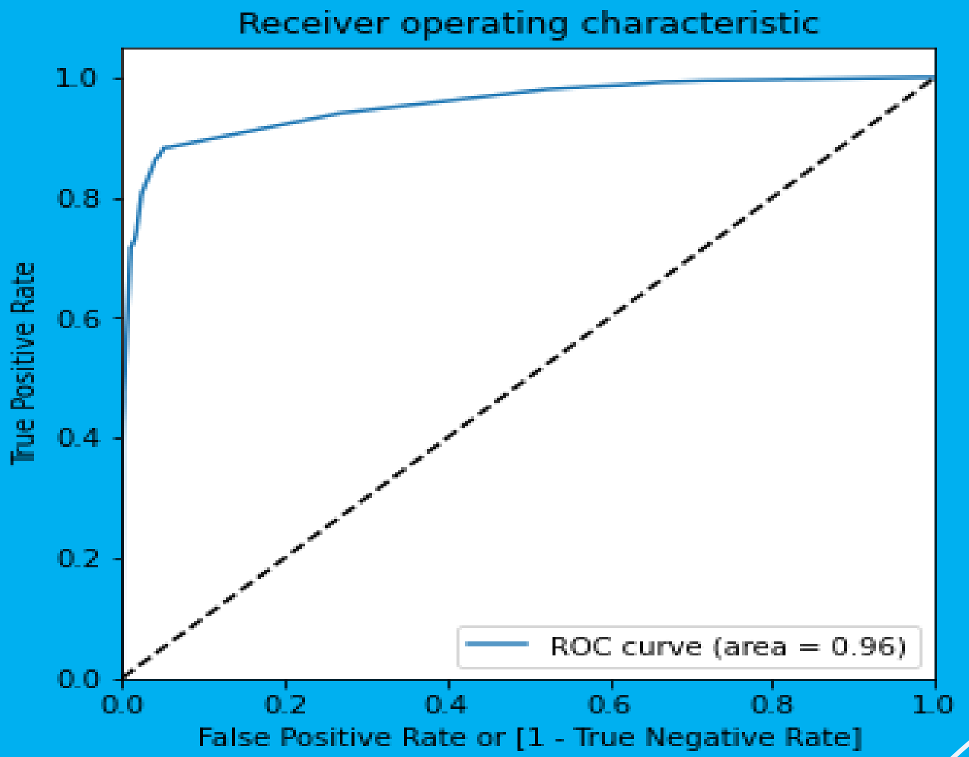
- Total visits
- Page views per visit
- Time spend on website



ROC CURVE

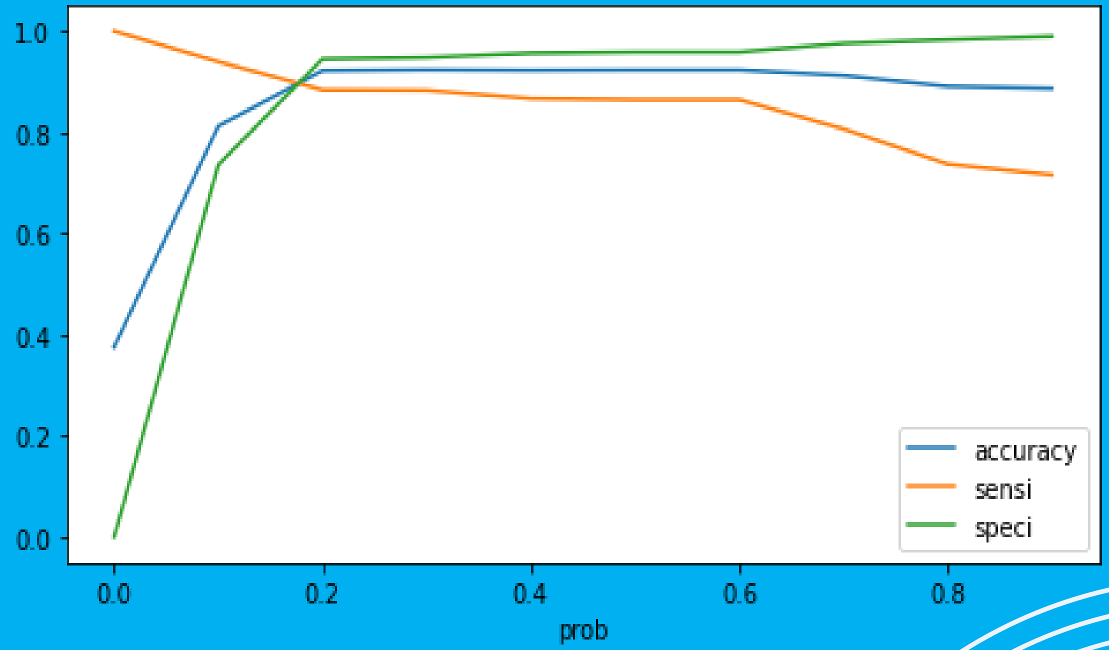
96% of the area is under ROC curve.

• Classification Probability of lead conversion by the model is very high.



Optimal Cutoff

Optimal Probability Cut-off With 0.2 cut-off, the model has: Accuracy – 92%
Sensitivity – 88%
Specificity – 94%



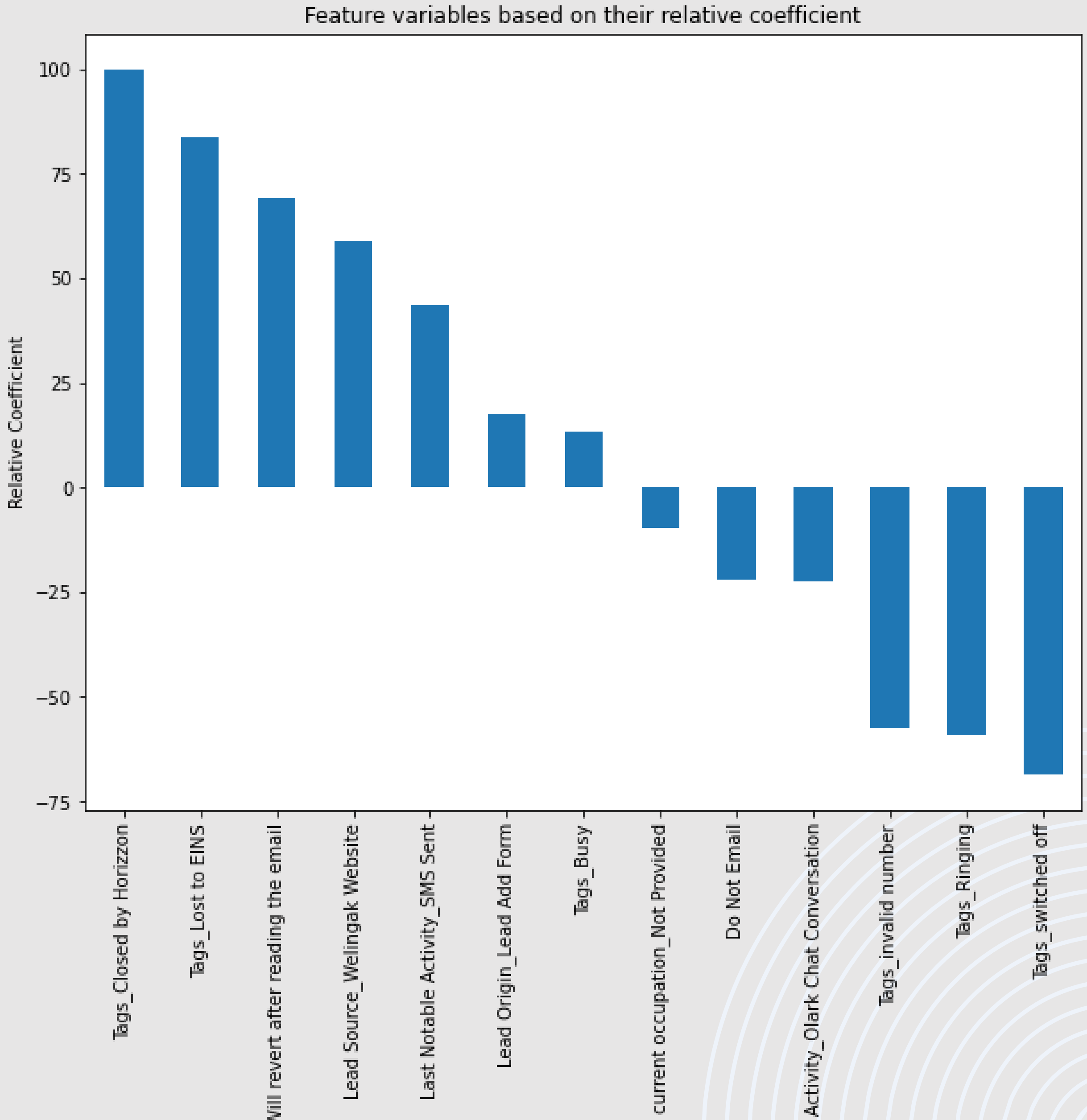
IMPORTANT FEATURES

Top 3 variables that have high lead conversion probability

- Tags – Closed by Horizzon, Lost to EINS, Will revert after reading the email
- Lead Source – Welingak Website
- Last Notable Activity – SMS Sent

Top 3 variables that need improvement in converting quality lead

- Tags – Invalid Number
- Tags – Ringing
- Tags – Switched off



Recommendations

- Follow ups through calls and emails with high conversion probability leads is suggested.
- Focus more on customers who spend a lot of time on the company's website as their conversion rate is high as per EDA.
- Providing special offers to customers who are highly interested and are seen visiting back to the website.
- Leads who have Tags such as 'Ringing', 'Switched Off', 'Invalid Number' can be avoided as the probability of them converting is very low

THANK YOU

-Satyam Khorgade