upGrad

International Institute of Information Technology Bangalore

Lead Scoring Case Study Using Logistic Regression

Identification of high converting lead

Shyam Mohan Azad
angelnikhilrohilla@gmail.com

Preeti Panda
preeti.p0016@gmail.com

Kaushik Pal
koushikbackin@gmail.com

Contents

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. They have process of form filling on their website after which the company identify that individual as a lead.

- 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, this means 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.

Objective

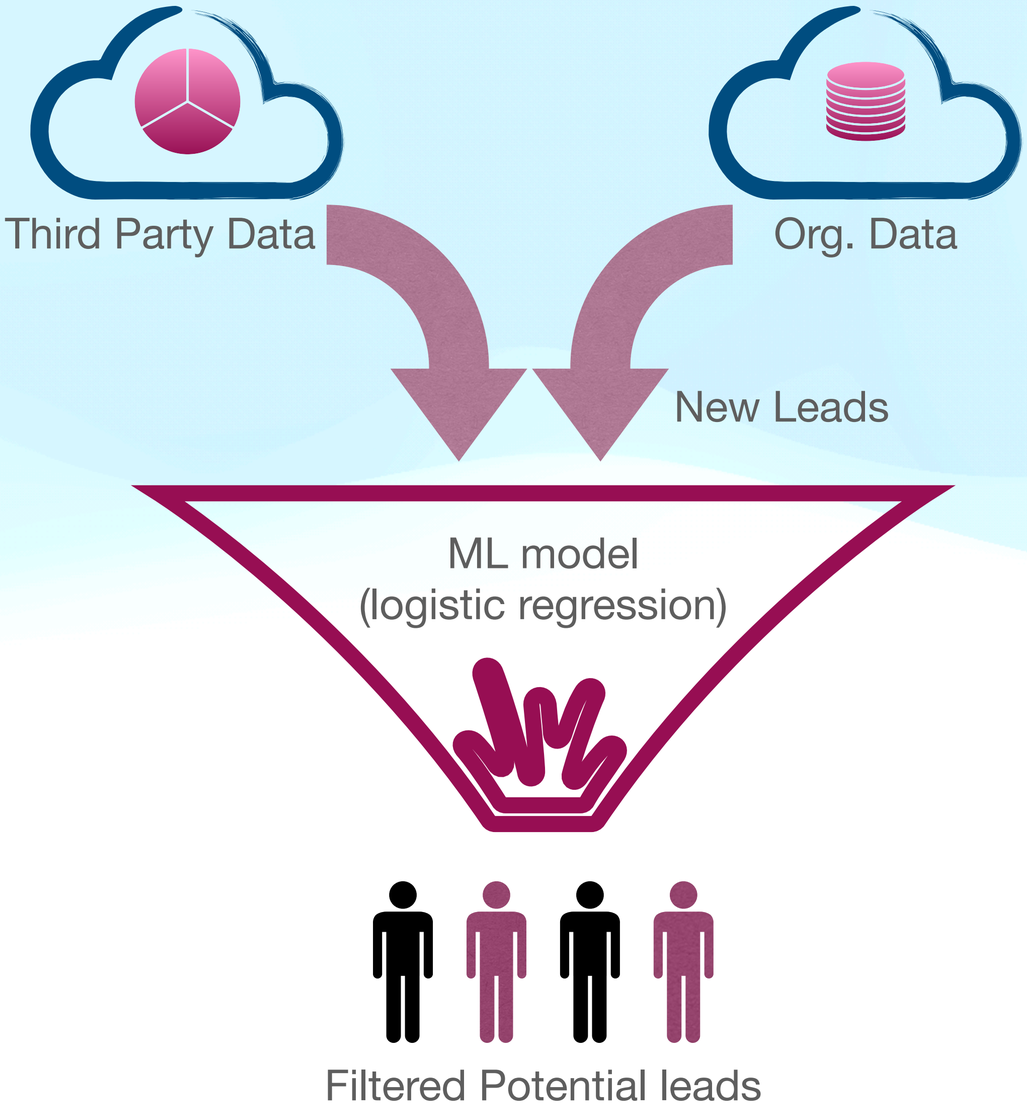
- Lead X wants us to build a model to give every lead a lead score between 0 -100 . So that they can identify the Hot leads and increase their conversion rate as well.

- The CEO want to achieve a lead conversion rate of 80%.

- They want the model to be able to handle future constraints as well like Peak time actions required, how to utilise full man power and after achieving target what should be the approaches.

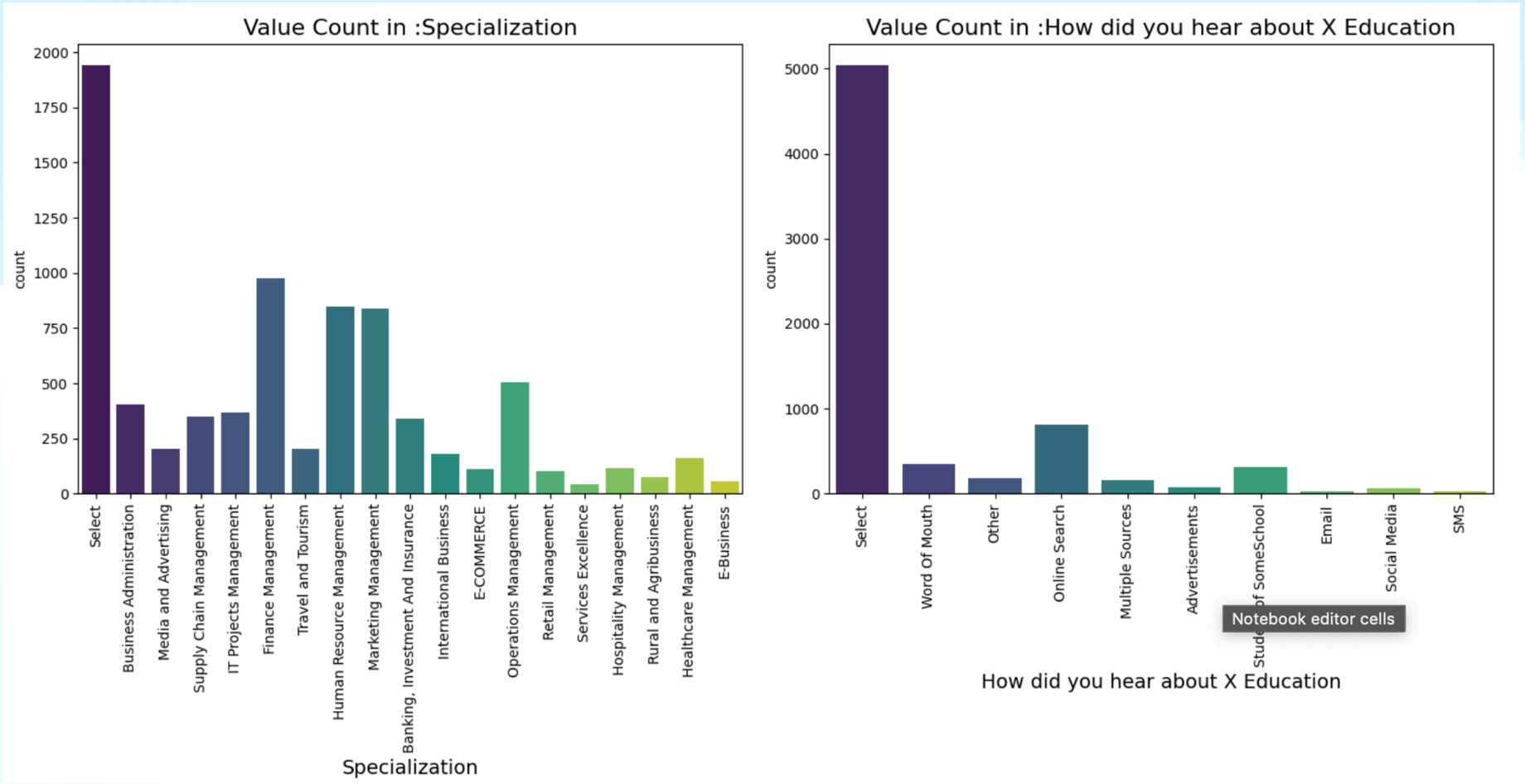
Problem Approach

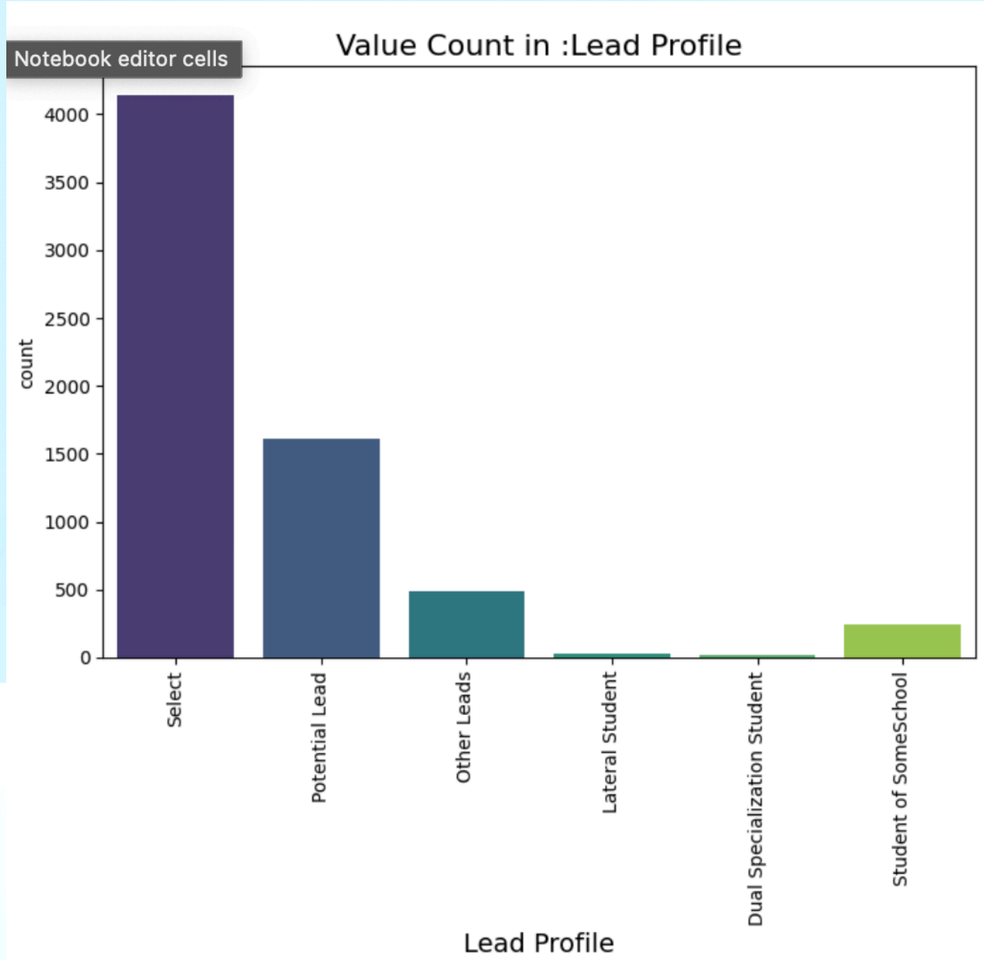
- Importing the data and inspecting the data frame
- Data preparation
- EDA
- Dummy variable creation
- Test-Train split
- Feature scaling
- Correlations
- Model Building (RFE Rsquared VIF and p- values)
- Model Evaluation
- Making predictions on test set

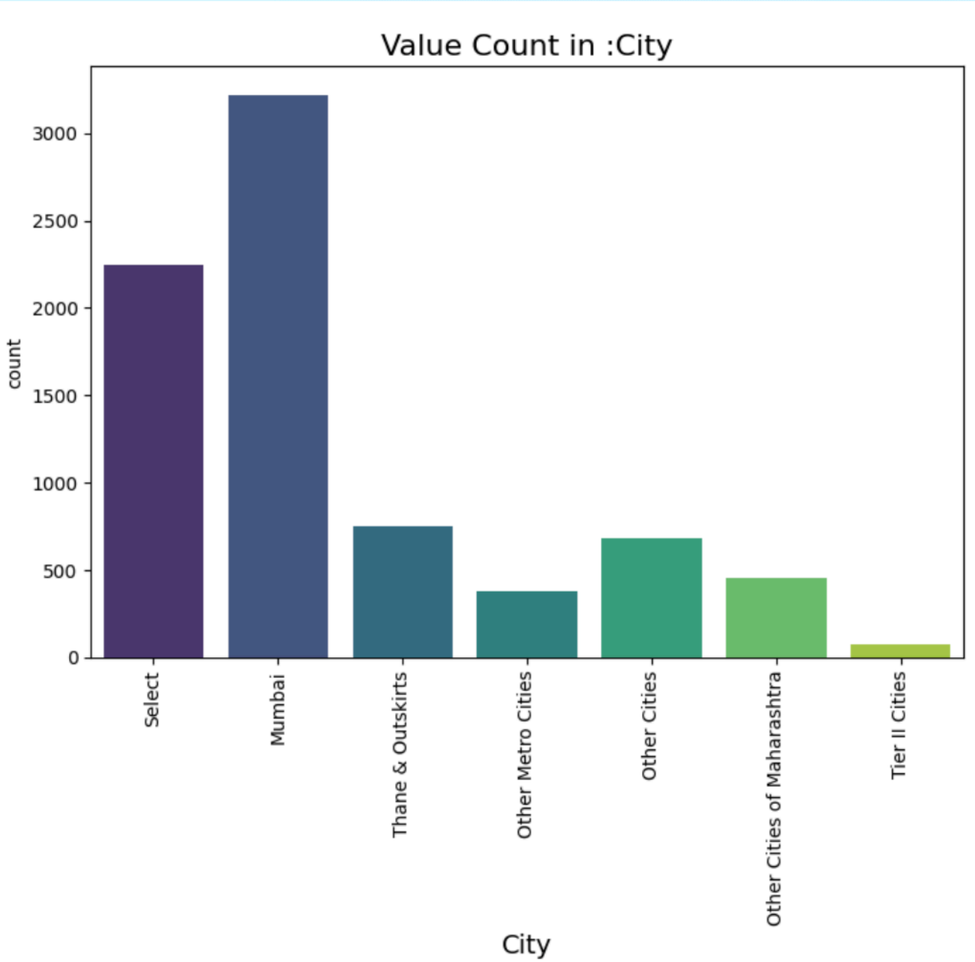


EDA - Data Cleaning and Preparation

All the nulls and a value “select” which is similar to missing value was treated.
We deleted columns(except Specialisation) having such values in each columns as they were of no importance in our analysis.



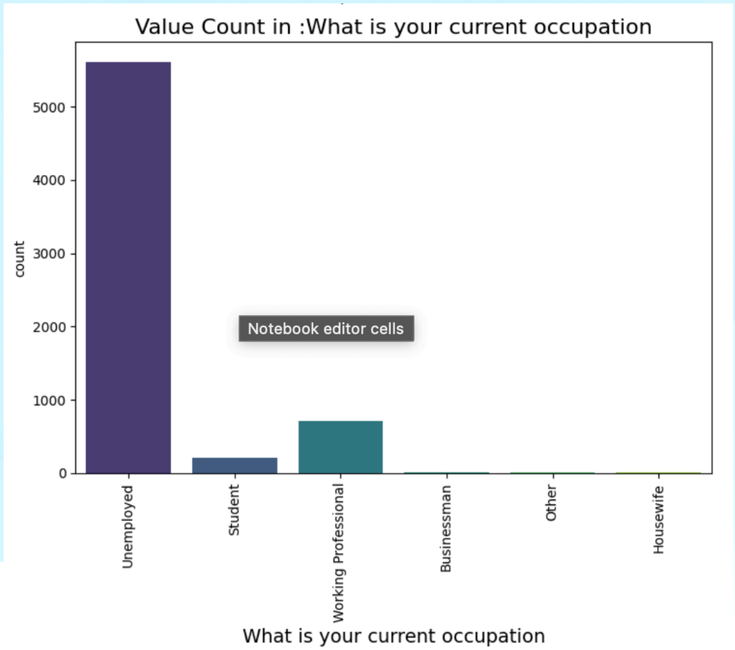


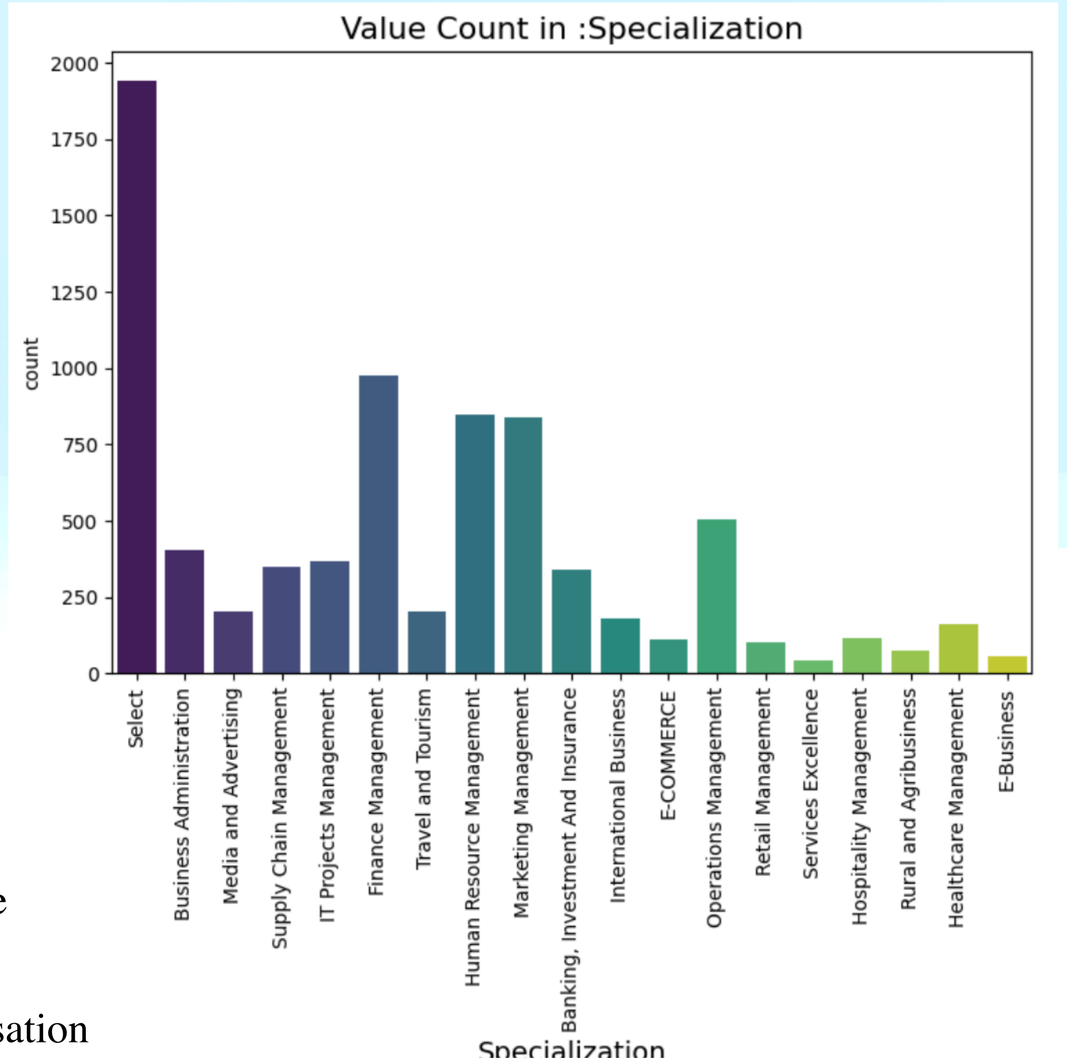


Why Column Specialisation not deleted ?

Though there was a very huge amount of value “Select” there, but we have other important information for various specialisation values which cannot be omitted without studying its impact on **lead conversion probability.**

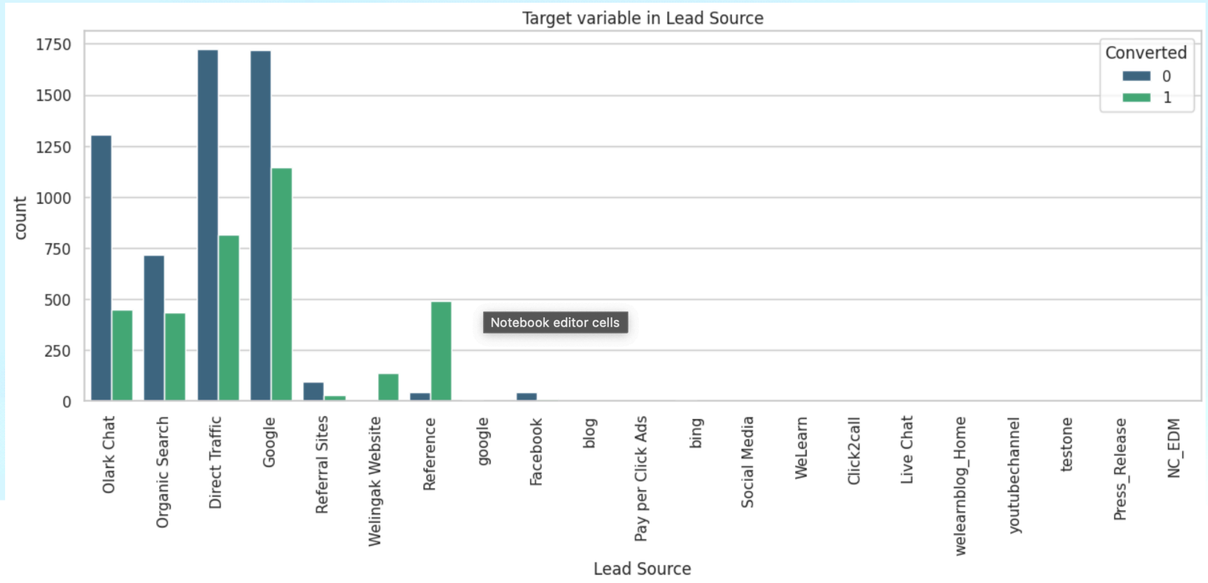
Specialisation

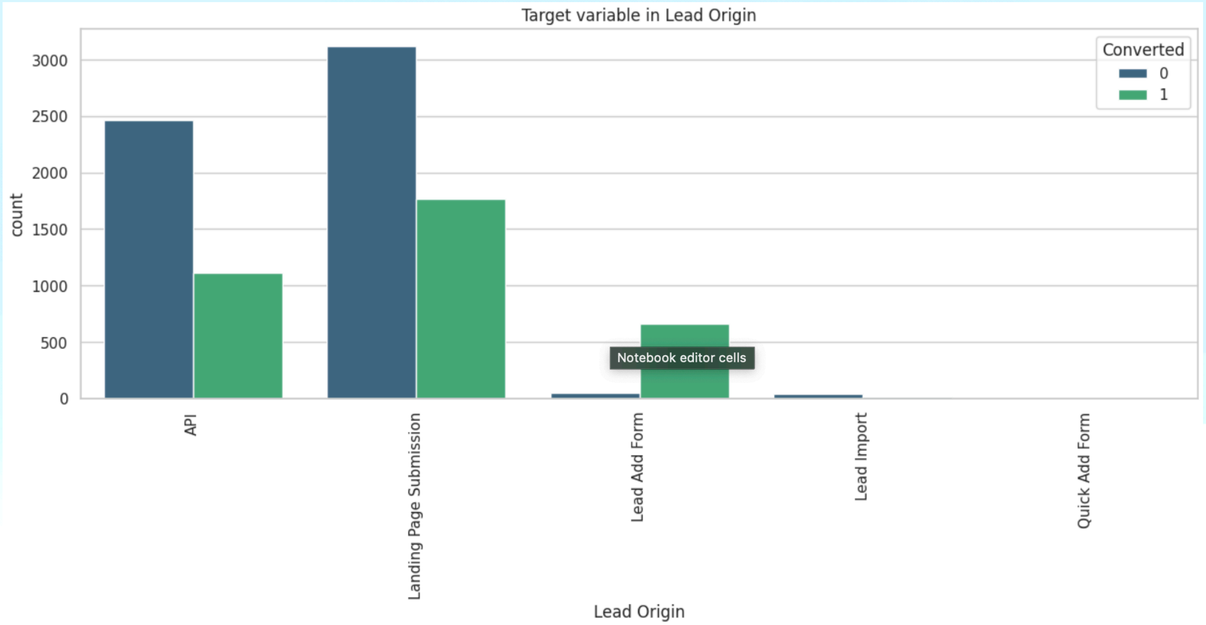




- It can be observed in the graph that majority of leads haven't chose any specialisation, reason may be we have observed that most of the leads are "Unemployed" as can be seen from "Current Occupation column.
- But we still have a good amount of information regarding lead's specialisation industry.
- It is observed that individuals from HR, Finance and Marketing Management fields are highly interested in Course.

Lead Source and Lead Origin



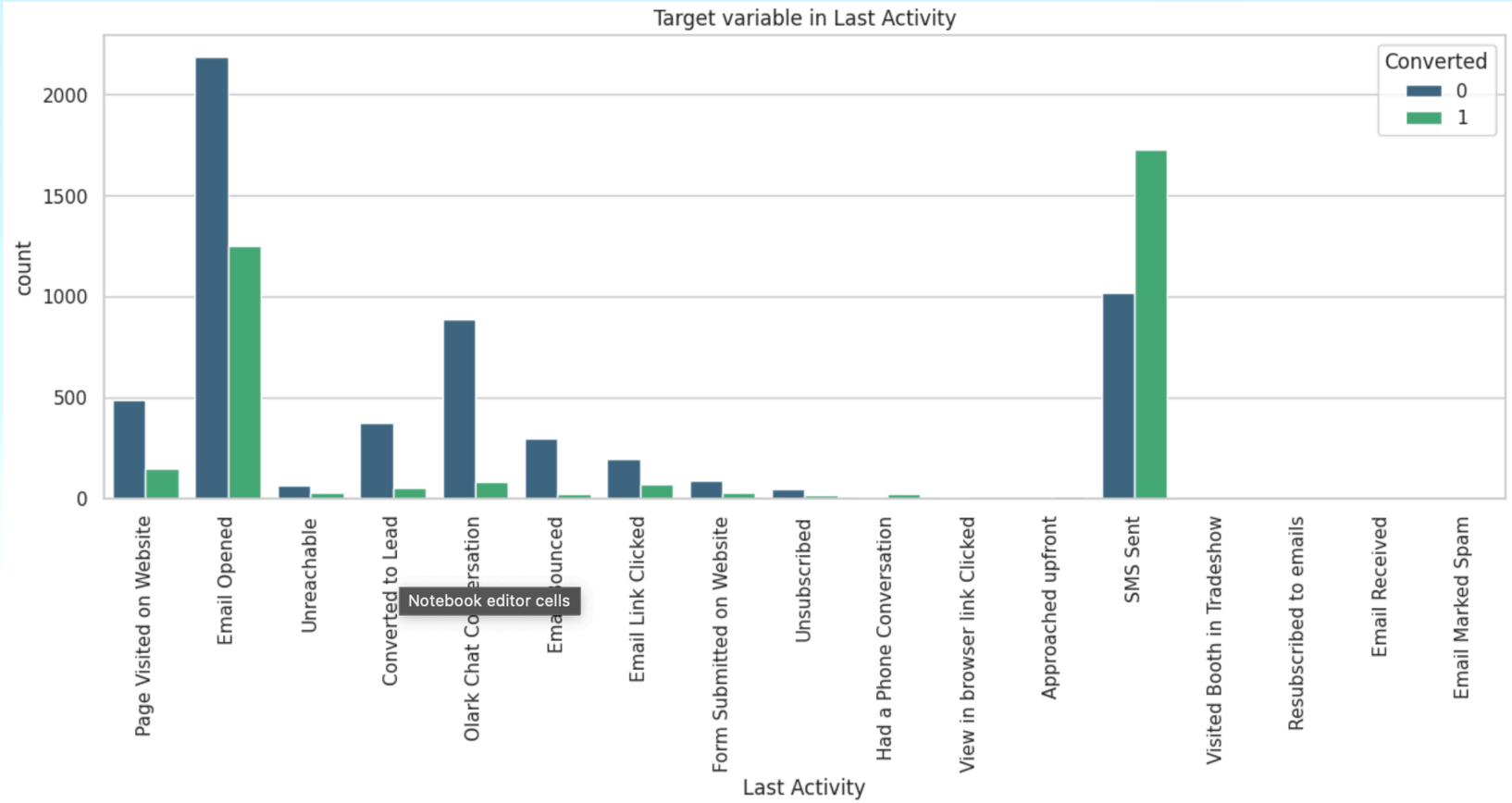


- In Lead Source it is observed people are more convertible from “Direct Traffic”, “Google Search” , “Olark chat” “and Organic Search”

- In lead Origin, landing page submission leads have highest share and then through API.

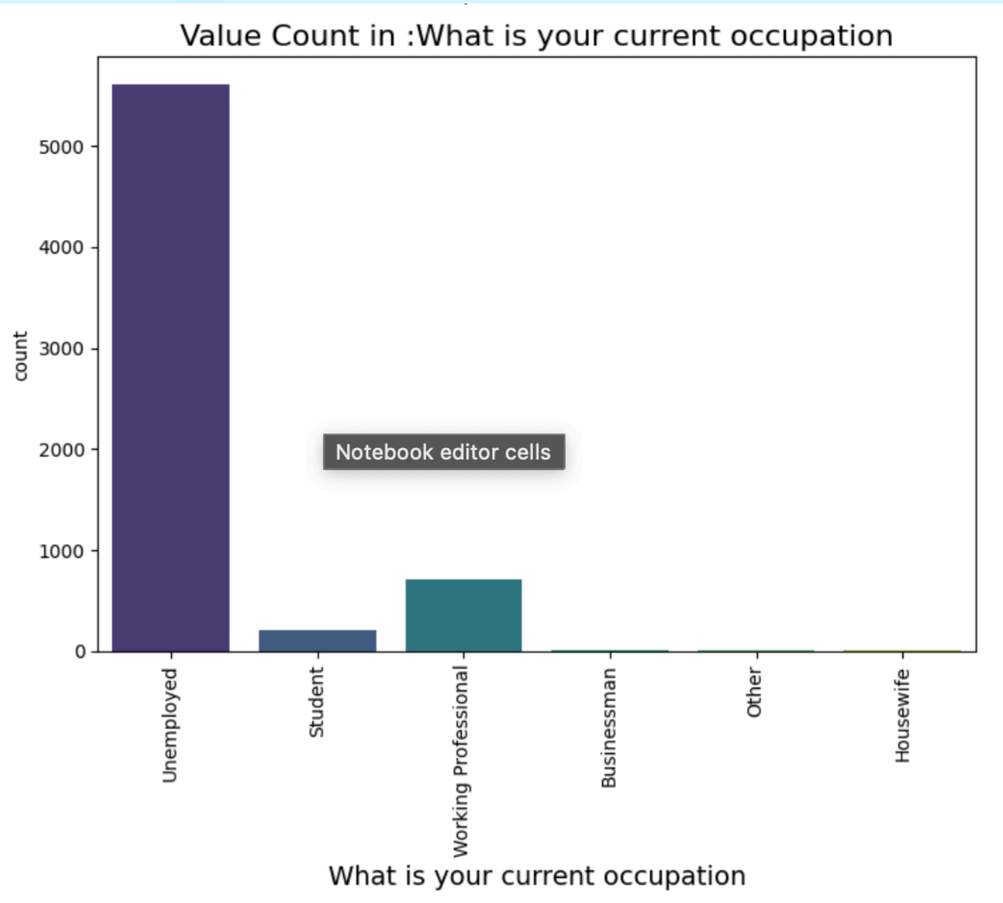
Last Lead Activity

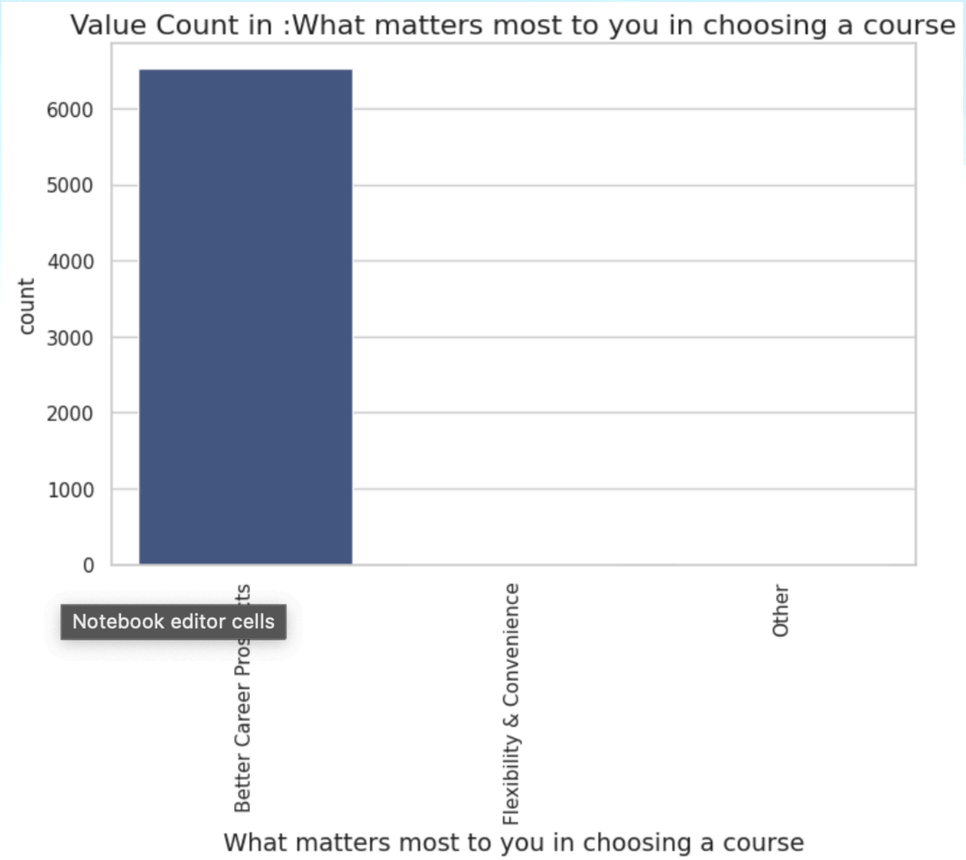
- Leads which are opening emails , having highest share in “Last Activity” column.
- A good amount of conversions also seen on SMS sent as well.



What is your Current Occupation

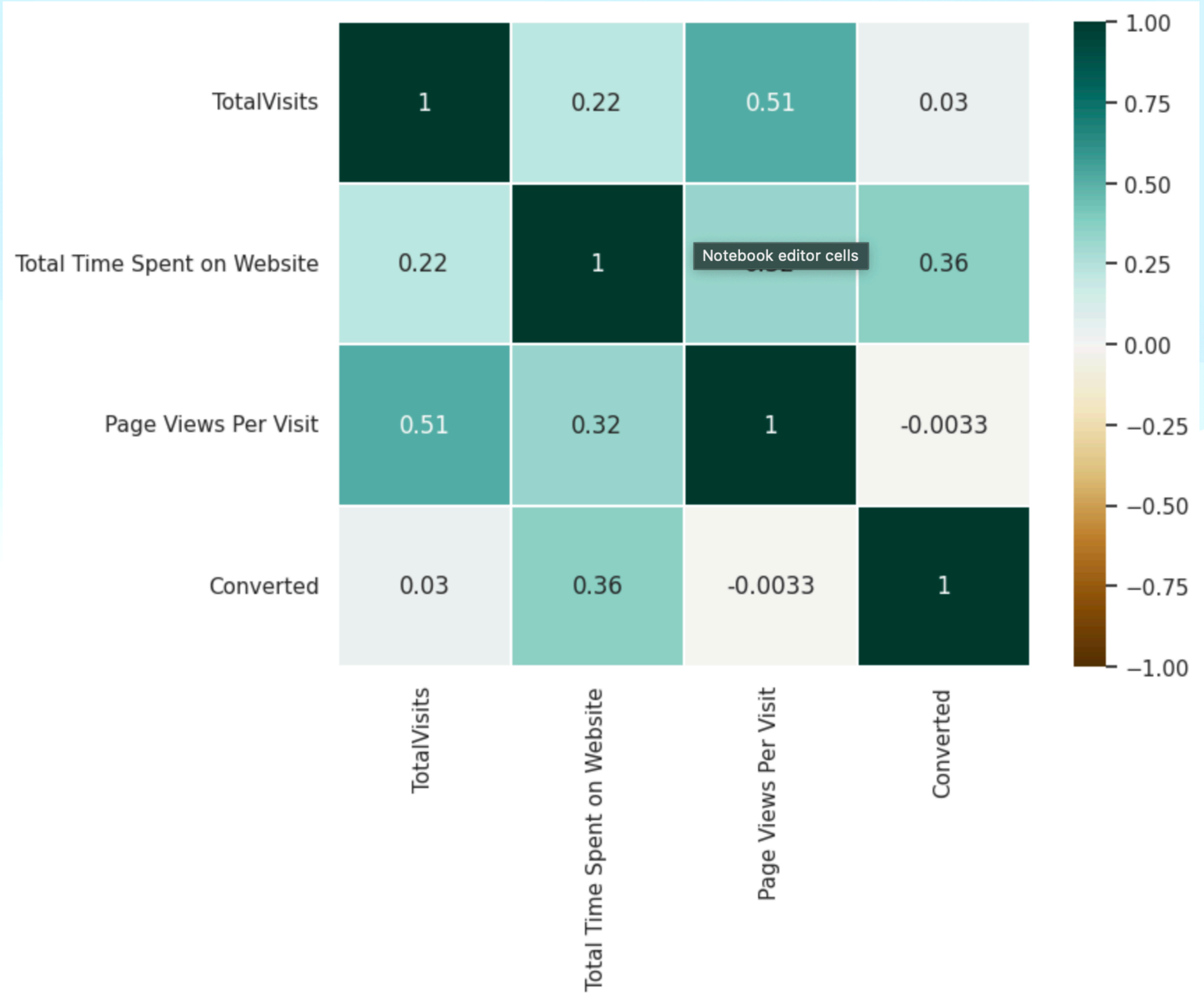
- More than 60% Unemployed, More than 27% Nulls.
- It is observed that Unemployed people are more interested in taking the course.
- Therefore we Dropped the column, as from here we can conclude most of the leads with high conversion rate and seeking a career start or transition





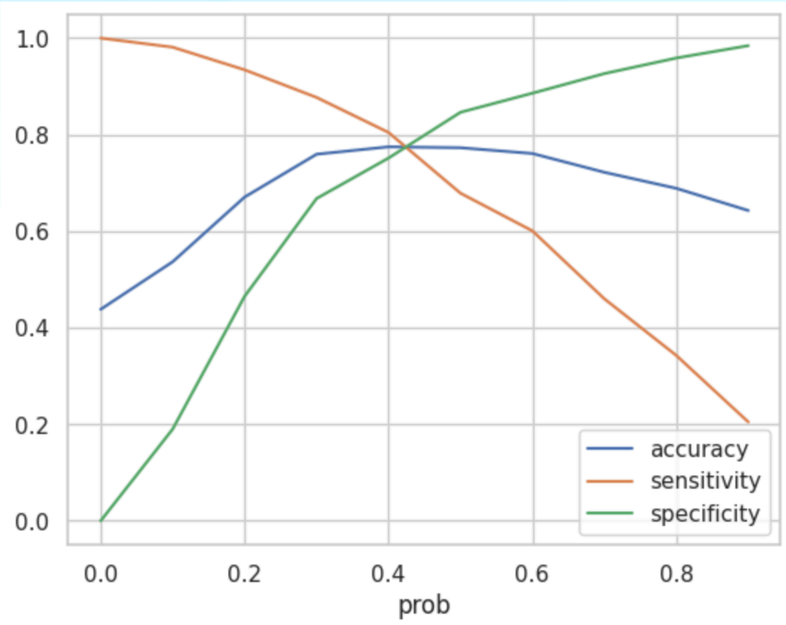
Correlation

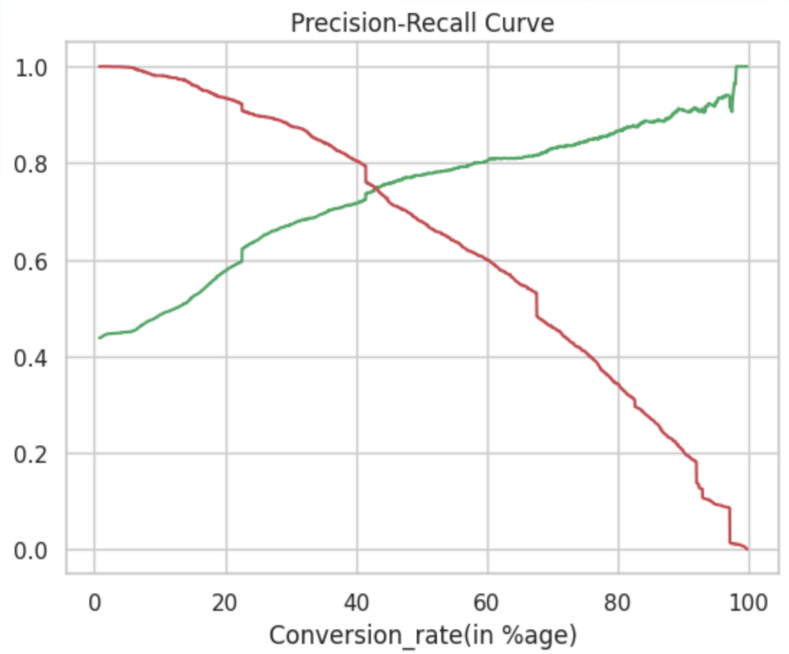
- No Significant correlation was found between variables.

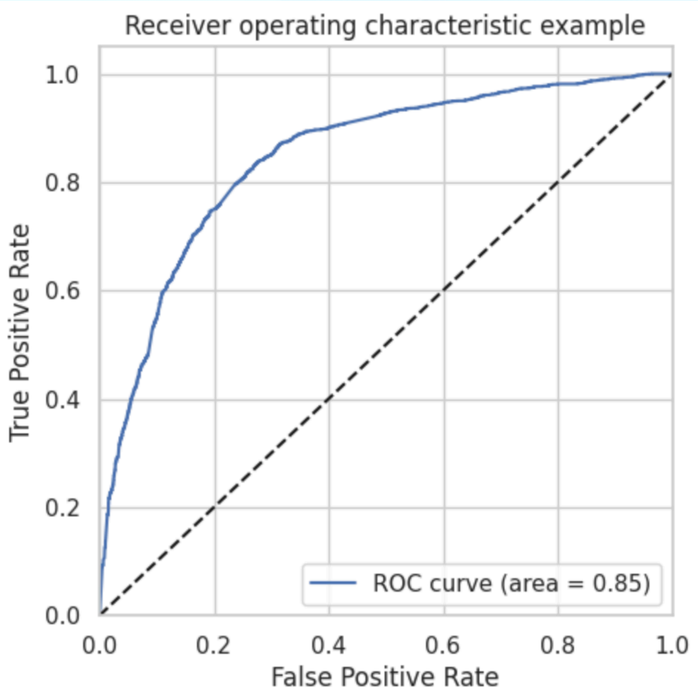


Model Evaluation

- **Optimum Cuttoff** observed **0.42 i.e 42%**
- **0.42** is the tradeoff between Precision and Recall -
  Thus we can safely choose to consider any Prospect Lead with Conversion Probability higher than **42 %** to be a hot Lead
- AUC in ROC curve was **0.85**







Observations

Training Data

Accuracy : 77.60%

Sensitivity: 75.68%

Specificity: 79.10%

Test Data

Accuracy : 76.88%

Sensitivity: 64.36%

Specificity: 85.58

Final Feature List

- ‘TotalVisits'
- 'Total Time Spent on Website’
- 'Page Views Per Visit’
- 'Lead Origin_Lead Add Form’
- 'Lead Source_Olark Chat’
- 'Last Activity_Email Bounced',
- 'Last Activity_Had a Phone Conversation’
- 'Last Activity_SMS Sent’
- 'Last Notable Activity_Email Opened’
- 'Last Notable Activity_Modified’
- 'Last Notable Activity_Olark Chat Conversation’
- 'Last Notable Activity_Unreachable'

Conclusion

- We observe that the conversion rate is around 30-35%, which is close to the average, for API and Landing Page submissions. However, the conversion rate is significantly lower for Lead Add forms and Lead imports. As a result, we can infer that it is essential to concentrate more on leads generated from API and Landing Page submissions.

- We notice that the highest number of leads are generated through Google/direct traffic. The maximum conversion ratio is observed from references and the Welingak website.

- Leads who spend more time on the website are more likely to convert..

- The most common last activity is email opened. The highest rate is for SMS sent. The maximum number are unemployed. The maximum conversion is with working professionals.