Lead Score Case Study

Case Study Group:

Archna Kottam

Ashutosh Shrivastava

Problem Statement

- X Education sells 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 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.

Solution Approach

- Problem Mapping
- Data Collection And Pre-processing
- Exploratory Data Analysis
- Model Building
- Model Evaluation

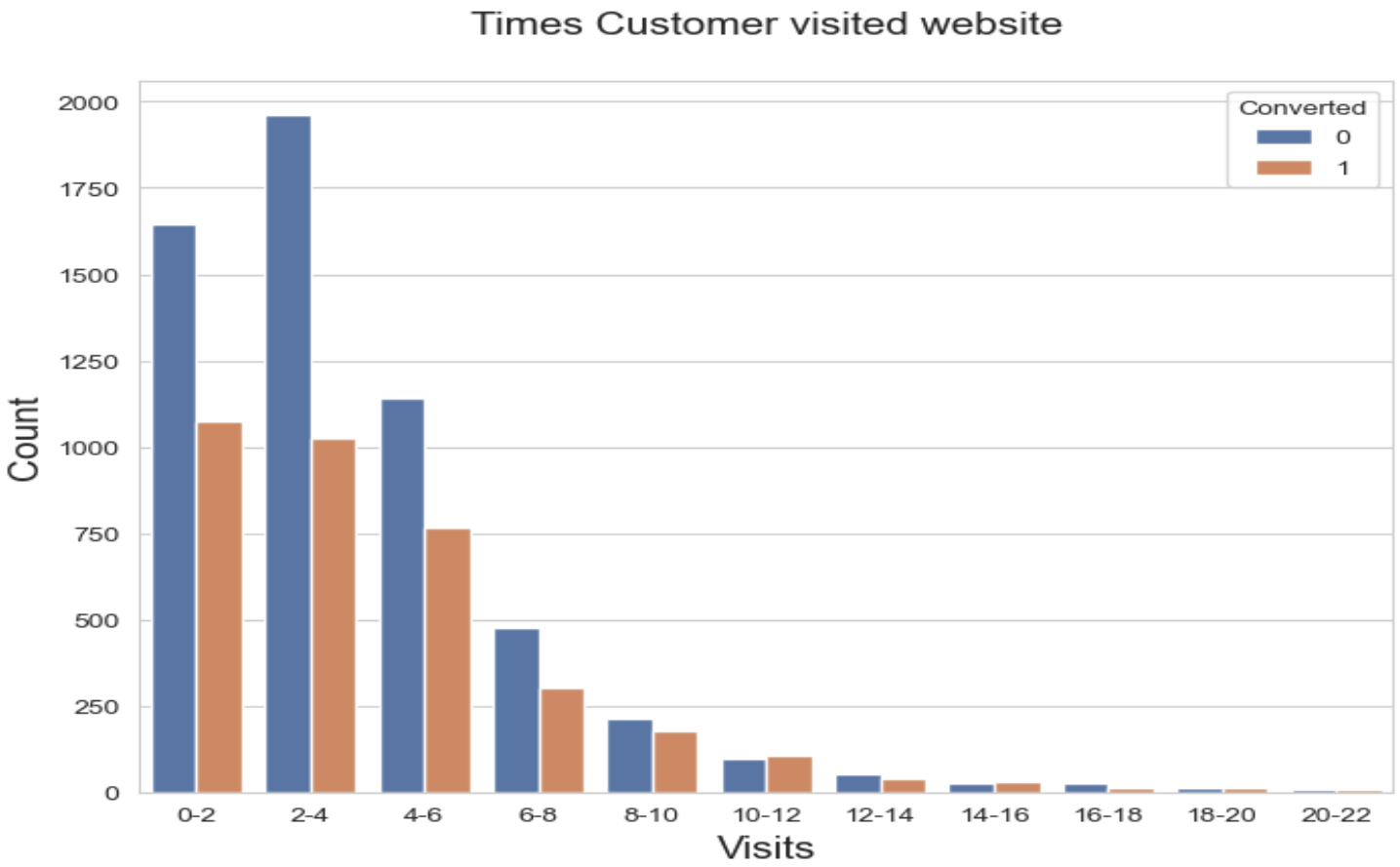
Problem Mapping

- X education wants to know most promising leads or Hot leads
- For that they want to a Model which can identifies the hot leads based upon the probability score for each leads.
- Deployment of the model for the future use so that their sales team could specifically target people with high chances of buying their products.
- On technical side the model should be such that it has high 'Recall' value i.e. for all the leads a company gets from different sources or medium, recall tells us how many got correctly identified or Hot leads.

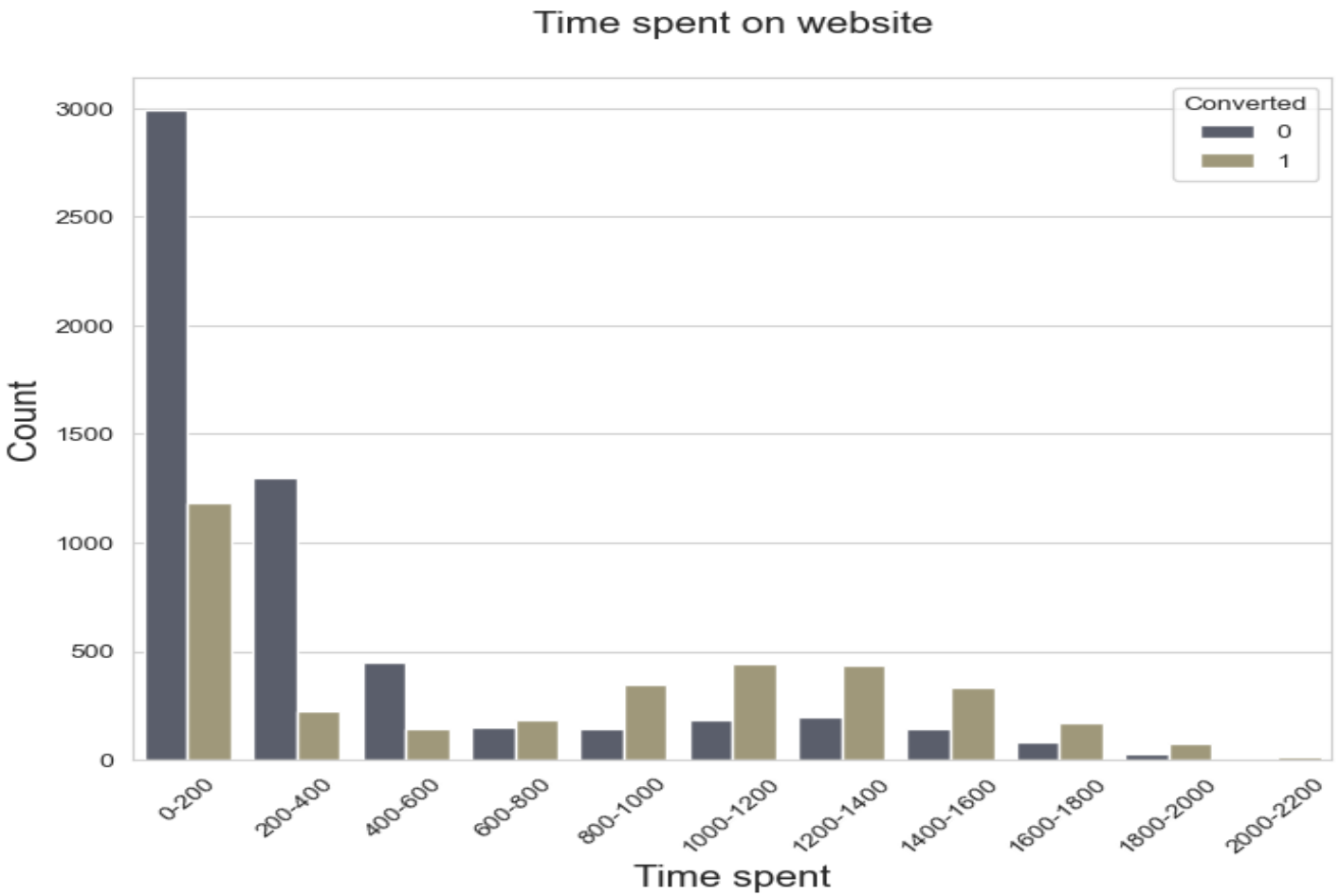
Data Collection and Pre-Processing

- Imported data from Leads.csv file.
- Familiarize with the data present in dataset.
- In few columns an unknown class ‘Select’ was present, which is nothing but the lack responses from customers. So converted this as null.
- Further check the percentage of null value and drop the columns with more than 40% values as null.
- Based upon data understanding further dropped some redundant columns to stabiles the complexity of model and saving the model from dimensionality curse.
- For the selected features imputed null values with another class ‘unknown’ and for continuous features imputed with mean or mode based upon feature relevance since these variables has some insights associated with them.
- Checked the imbalance ratio for our target variable ‘Converted’ which came to be 1.59 i.e. moderate.
- By performing these operations, we got our final feature variables(10 features) for further analysis.

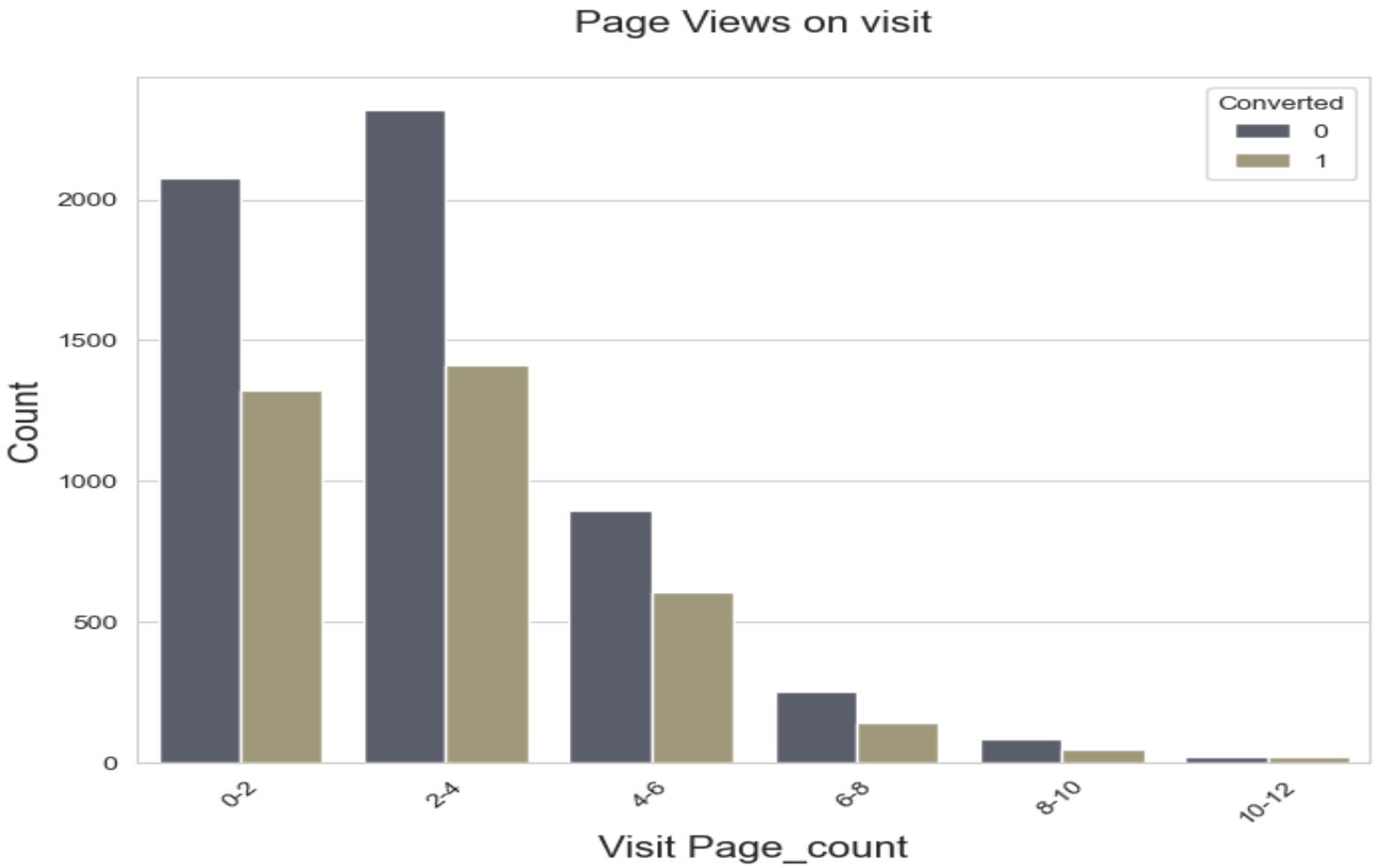
Exploratory Data Analysis



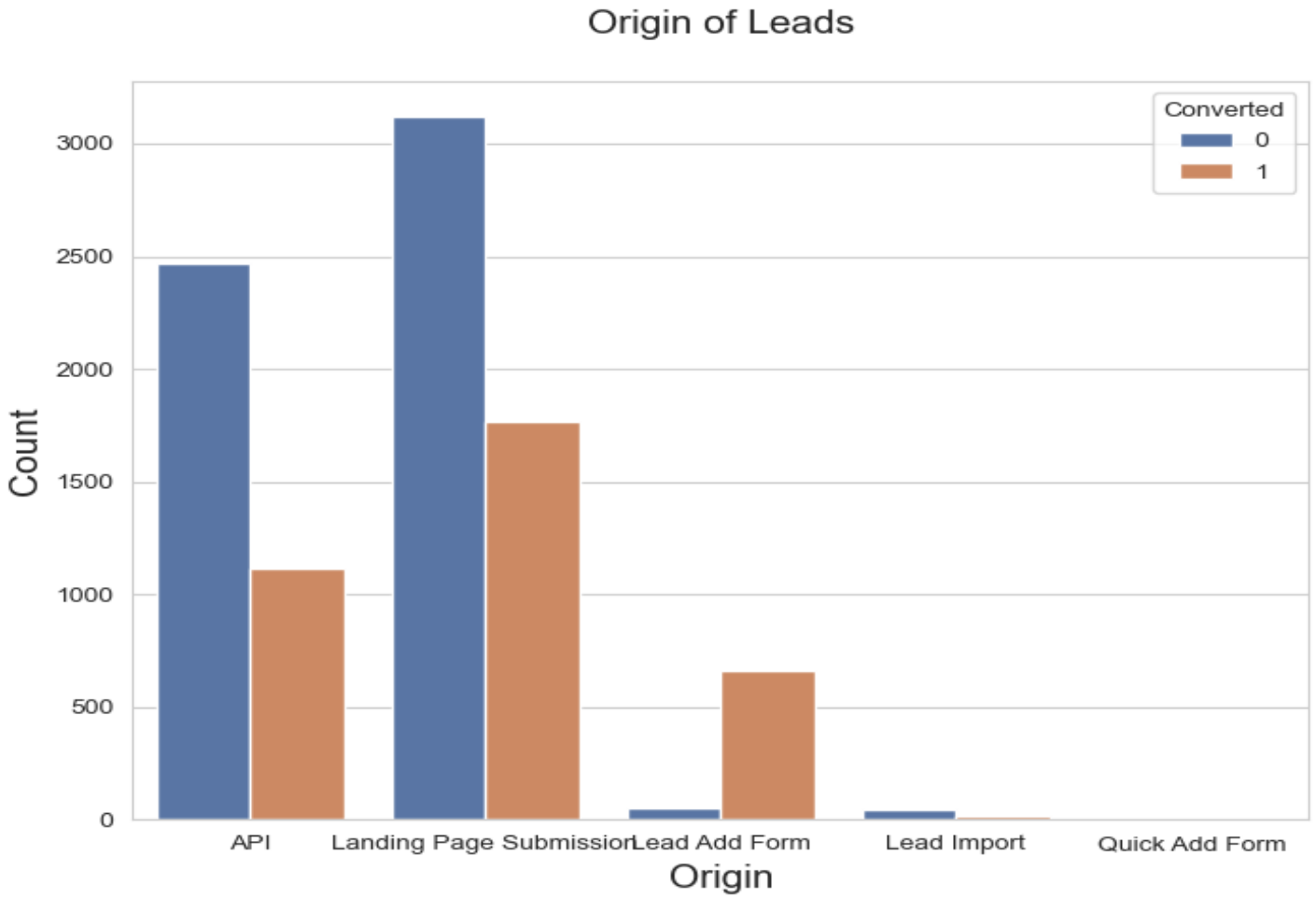
By binning the continuous variable, it shows that most of the conversion has happened for leads who visited website between 0 to 6 times. Which signifies that people who purchase product visited website for targeted search of courses available rather than browsing different available courses.



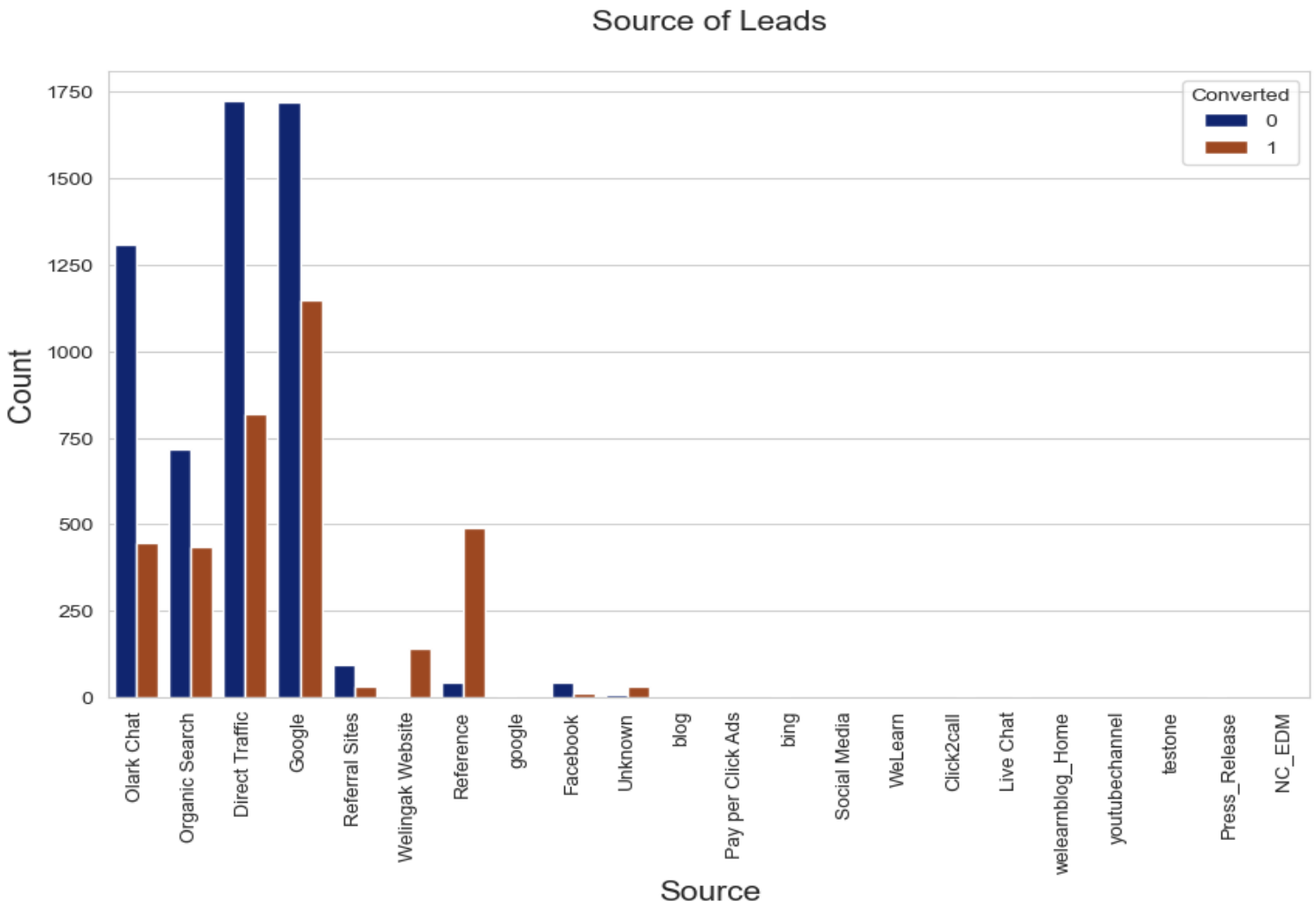
- Here also most of conversions have happened for leads who knew what they were looking for since their time spent on website is least comparing to others. Therefore, people with clearer mindset and have searched specific product instead of browsing multiple products are potential leads.



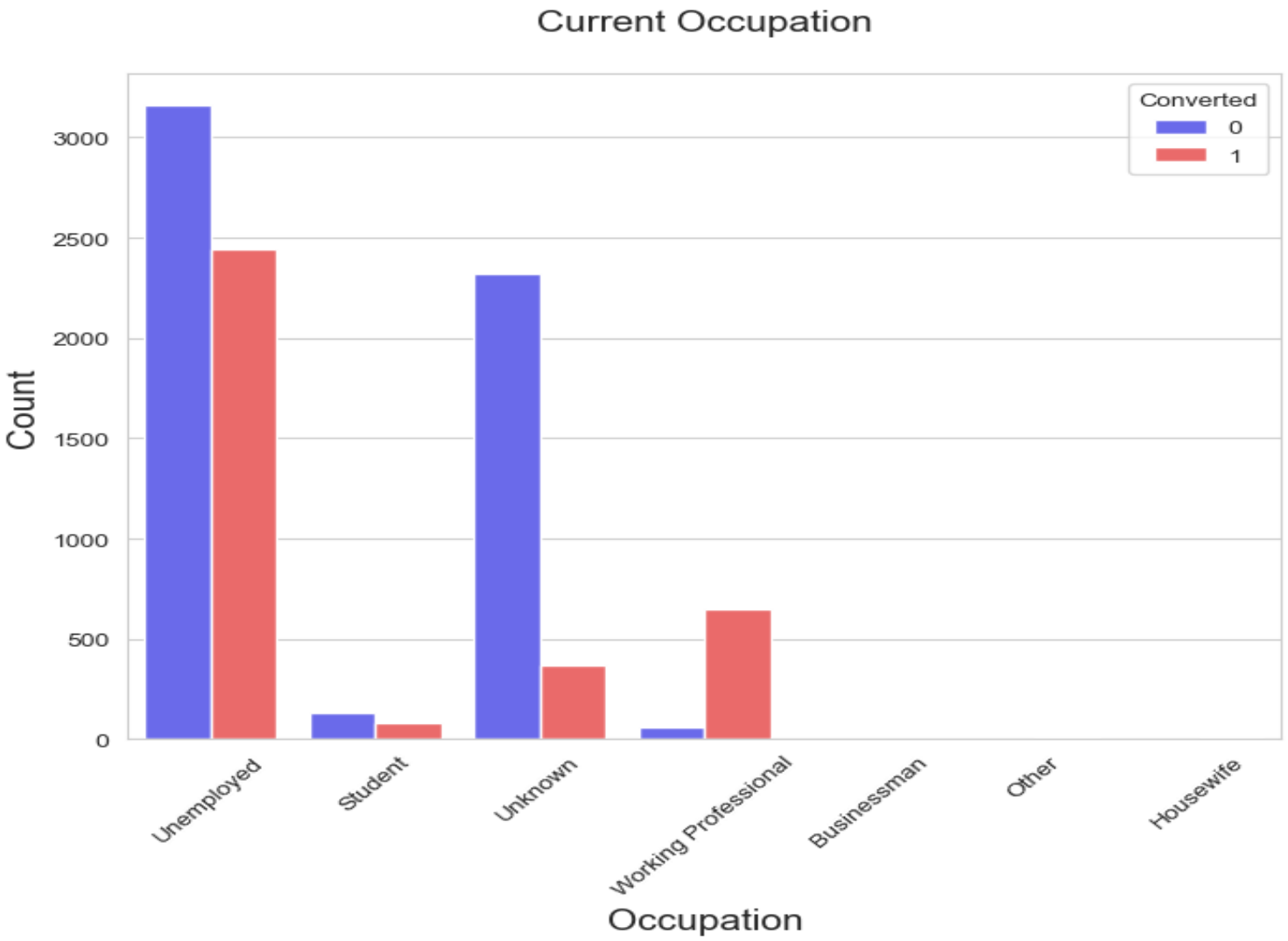
- Conversion is higher among leads with view count of pages between 0 to 4 which justifies our earlier analysis i.e., leads with specific product search are hot leads.



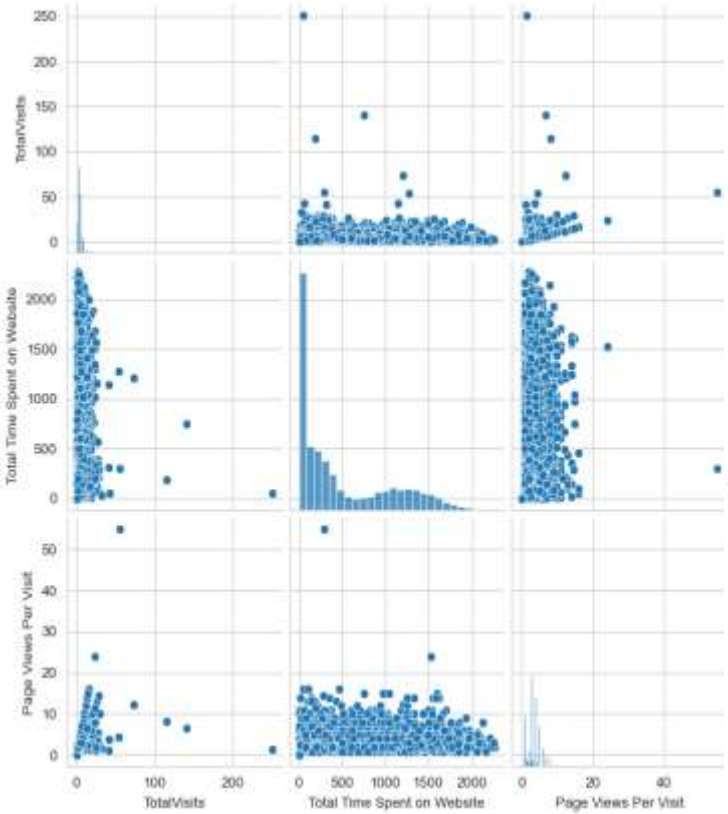
- 'Lead Add form' is a clear winner with highest conversion ratio whereas 'API' and 'Landing Page Submission' also showing good conversion in absolute term.

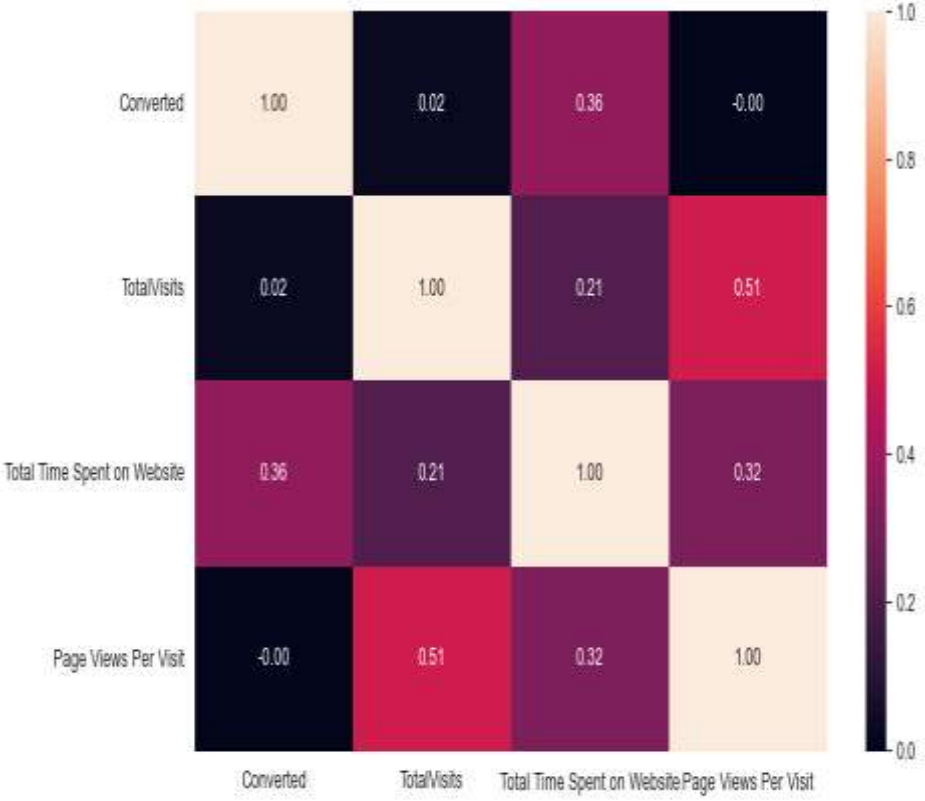


- 'Welingak website', 'Reference' have greater conversion ratio and in absolute value. 'Google' & 'Direct Traffic' have better conversion. 'Organic Search' show better trade off than 'Olark Chat'.



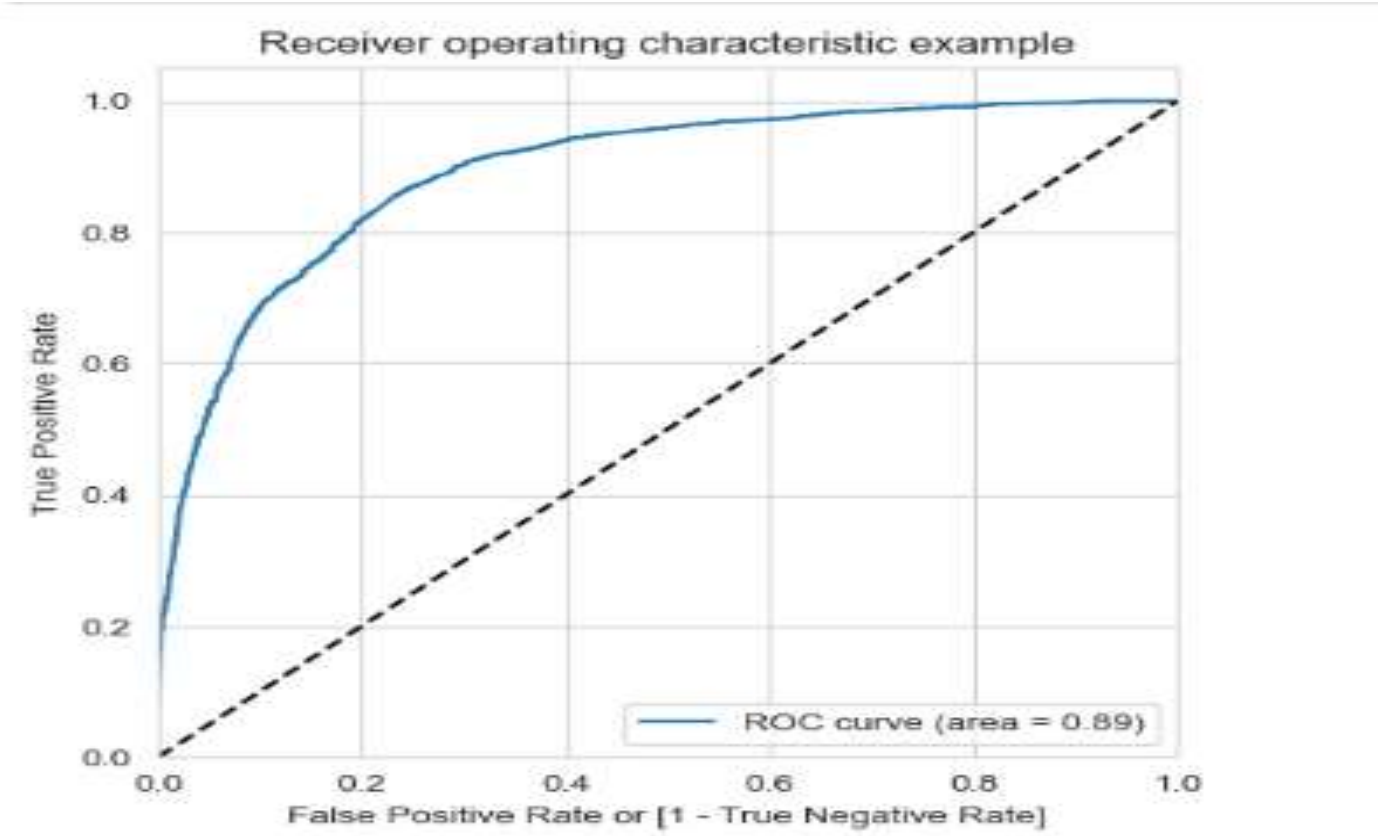
Unemployed and Working Professionals have better turn around. Special focus needs to be given to 'Working Professionals' as they have a very high conversion with fewer refusals.



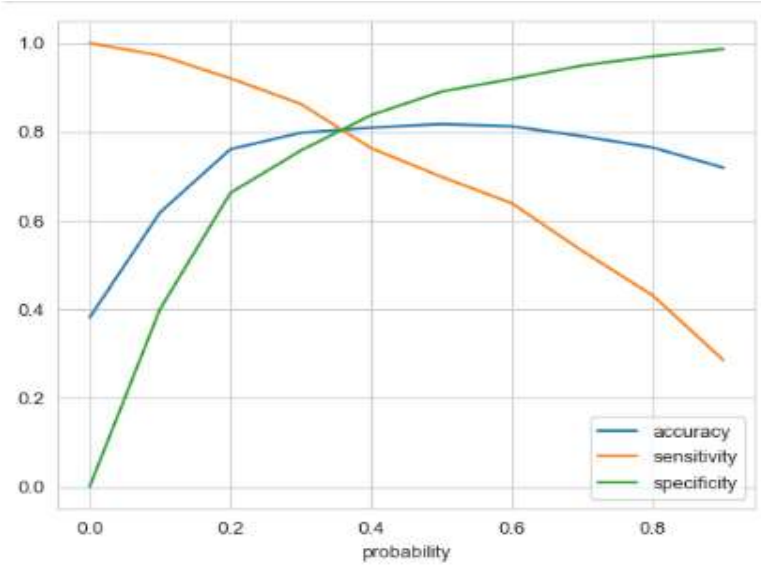


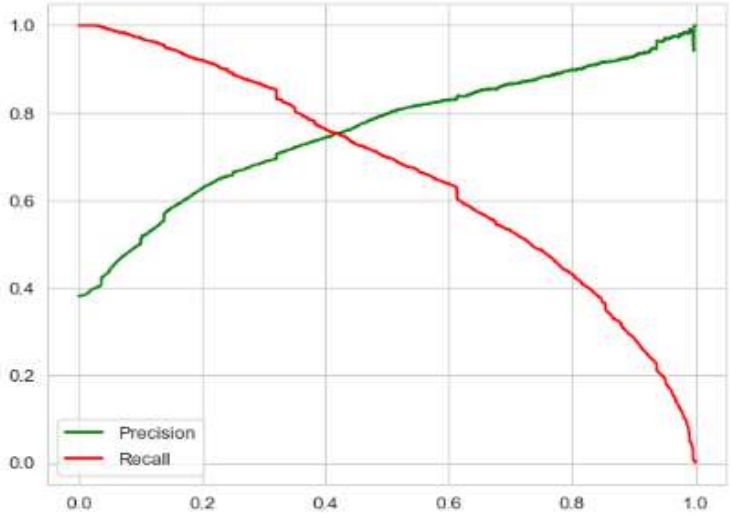
- There are no strong correlation present in data set we are good to start data preparation for modelling. Very few outliers are present so will keep it as it is since it won't affect our analysis!!

Model evaluation



The area under the curve of the ROC is 0.89 which is quite good. So we seem to have a good model.





0.35 is the tradeoff between Precision and Recall -
Thus, we can safely choose to consider any Prospect Lead with Conversion Probability higher than 35 % to be a hot Lead.

Conclusion

- It was found that the variables that mattered the most in the potential buyers are :
  1. The total time spend on the Website.
  2. Total number of visits.
  3. When the lead source was:
     a. Google
     b. Direct traffic
     c. Organic search
     d. Welingak website
  4. When the last activity was:
     a. SMS or Email opened
     b. Olark chat conversation
  5. When the lead origin is Lead add format.
  6. When their current occupation is as a working professional or unemployed.

  Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.