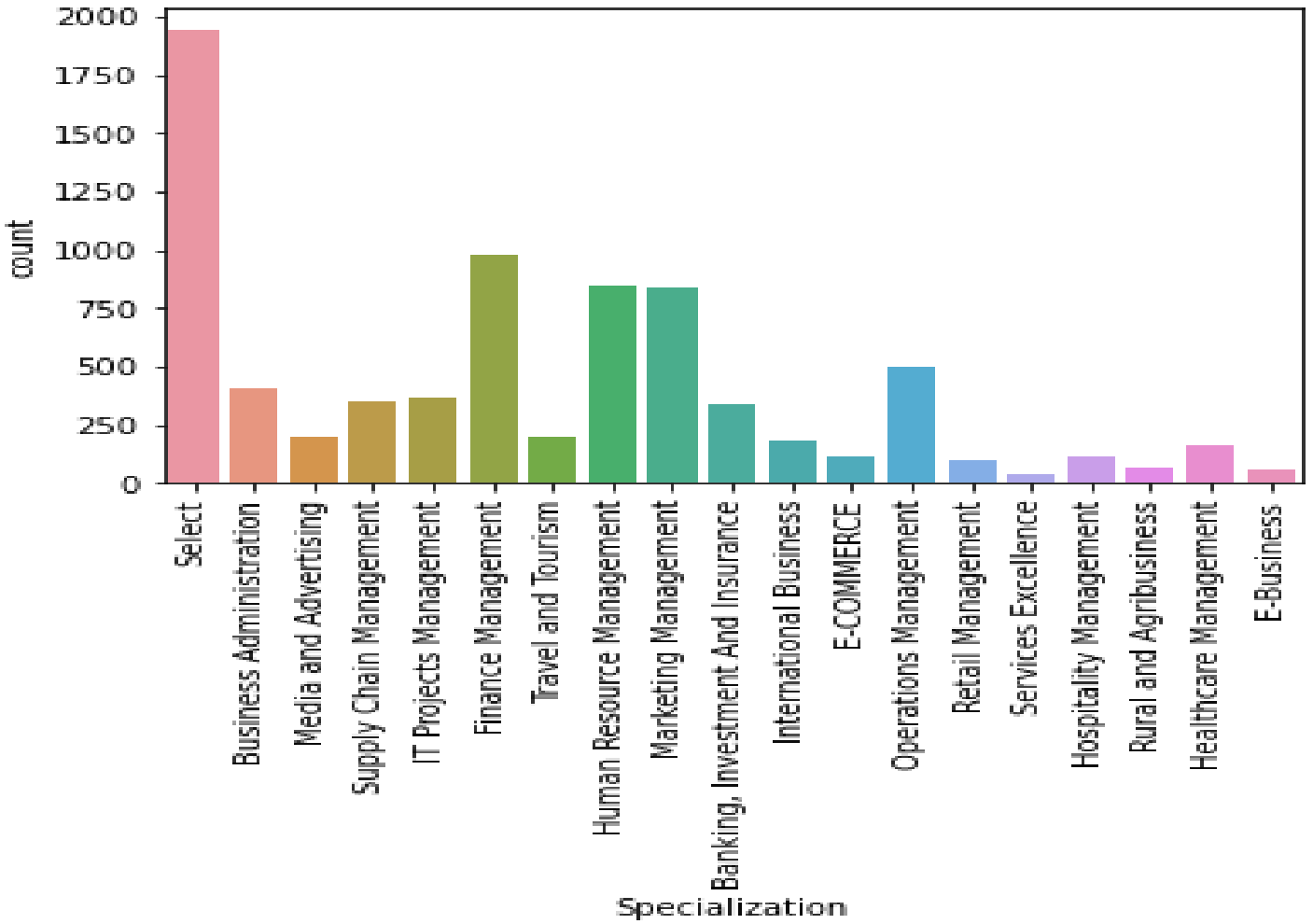
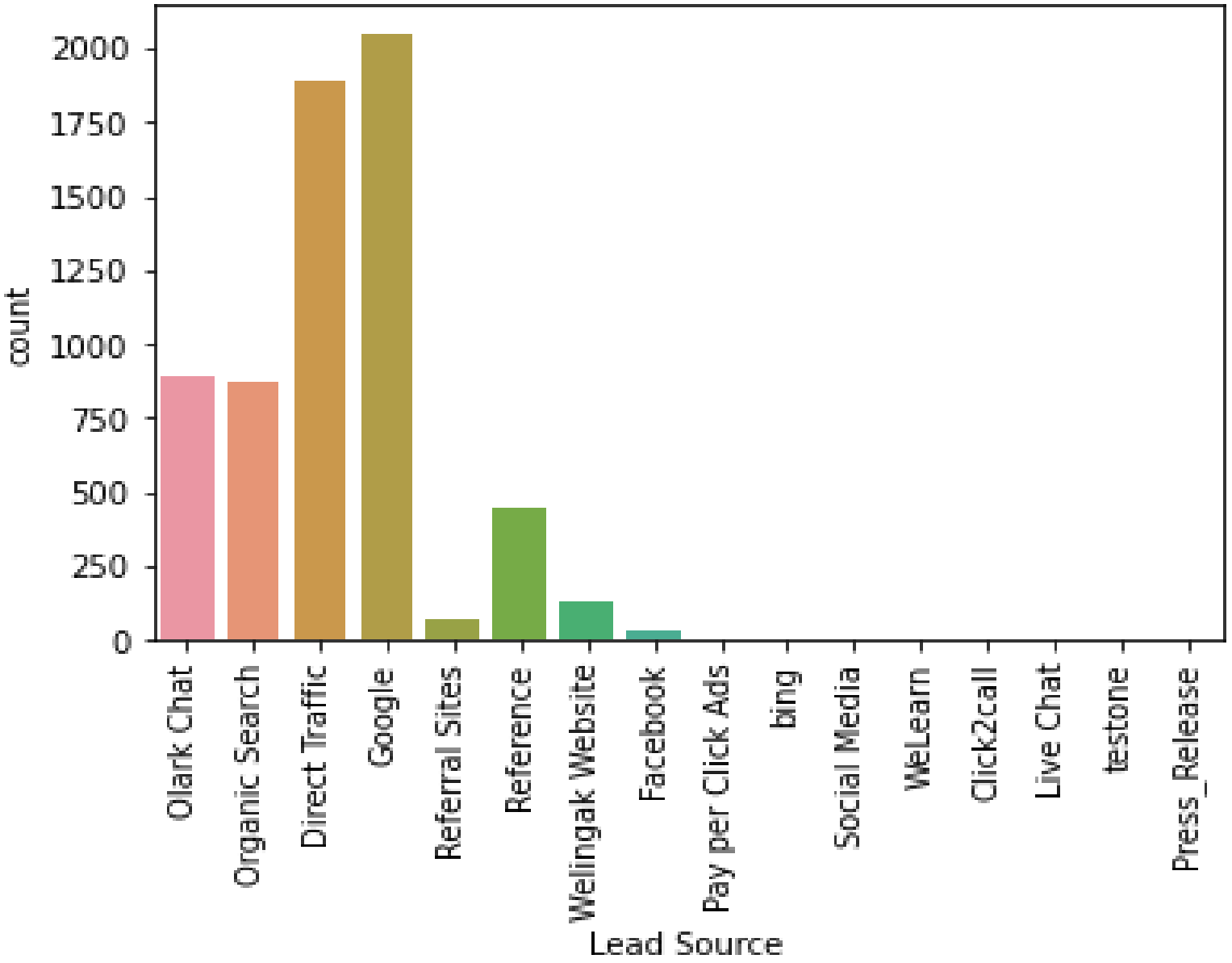
Lead Scoring Case Study

BY
Bobba Deerthika Shalini
Nachiketa
Richa Madan

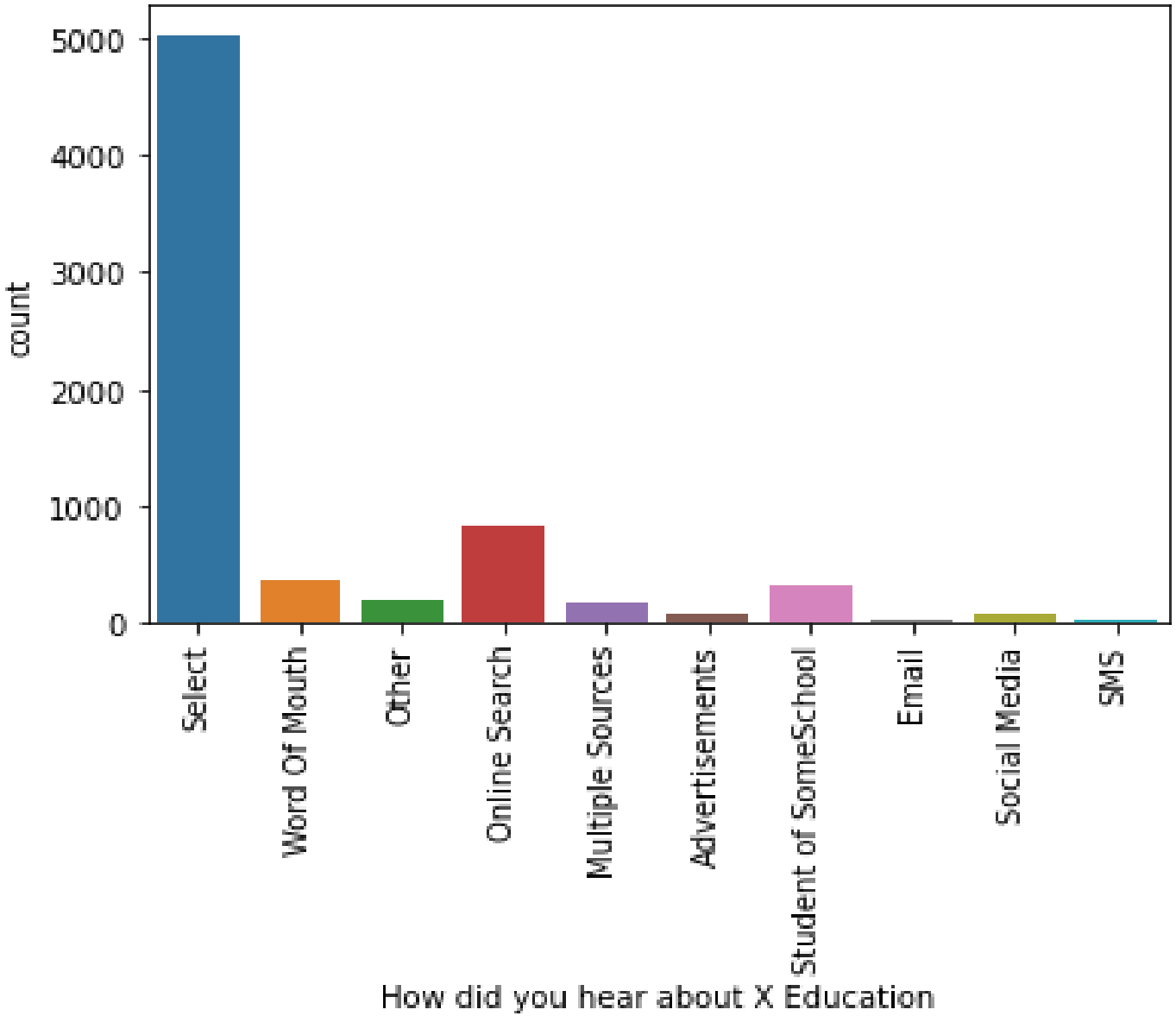
EDA



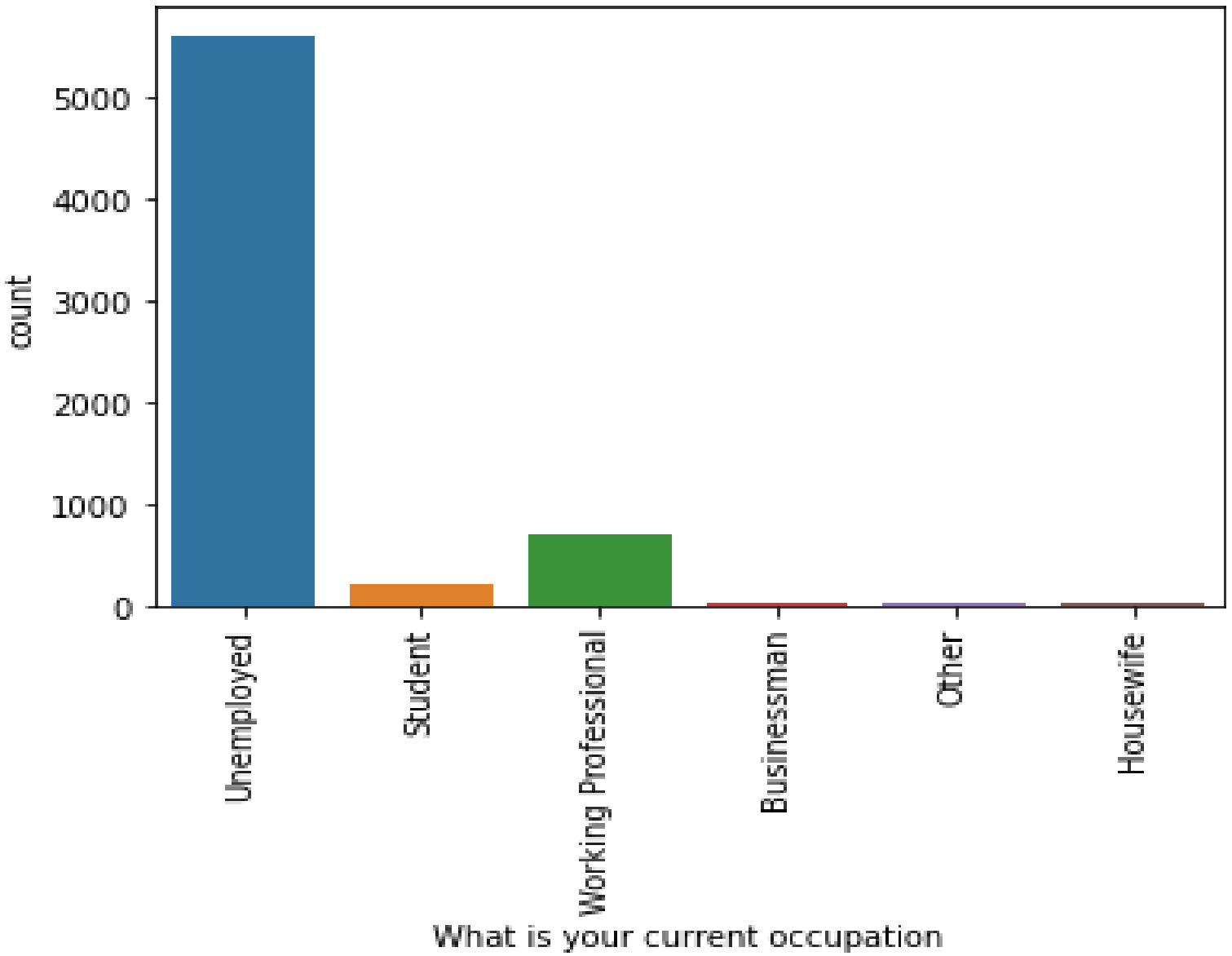
- Looking at above plot, no particular inference can be made for Specialization as select column is highest.
- Finance Management is second highest.
- HR Management, Marketing Management is third highest .
- Services Excellence is lowest among all.



- The count of leads from the Google and Direct Traffic is maximum.
- The conversion rate of the leads from Reference and Welingak Website is less compare to search.

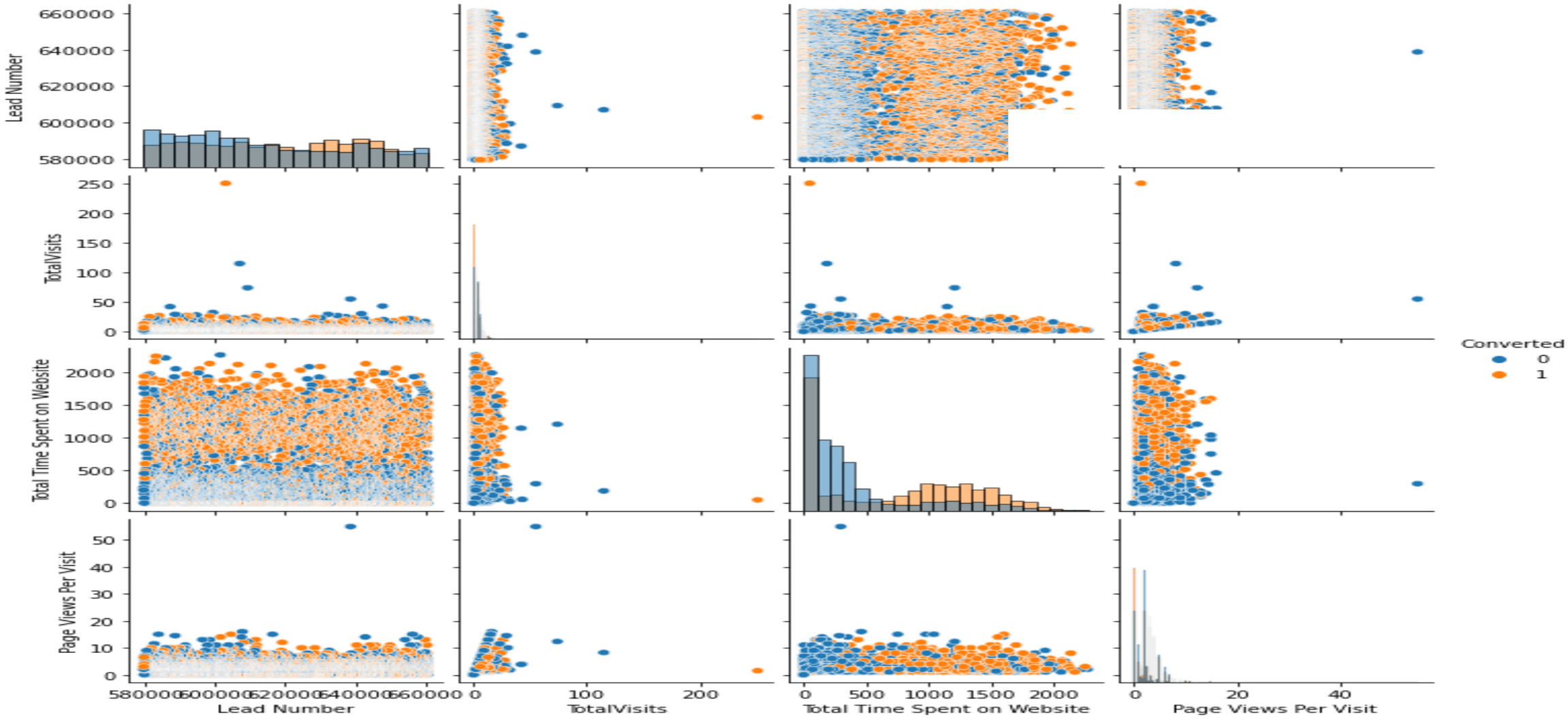


- Looking at above plot, no particular inference can be made for How did you hear about X Education as select column is highest.
- Online search is second highest.
- SMS is lowest among all.

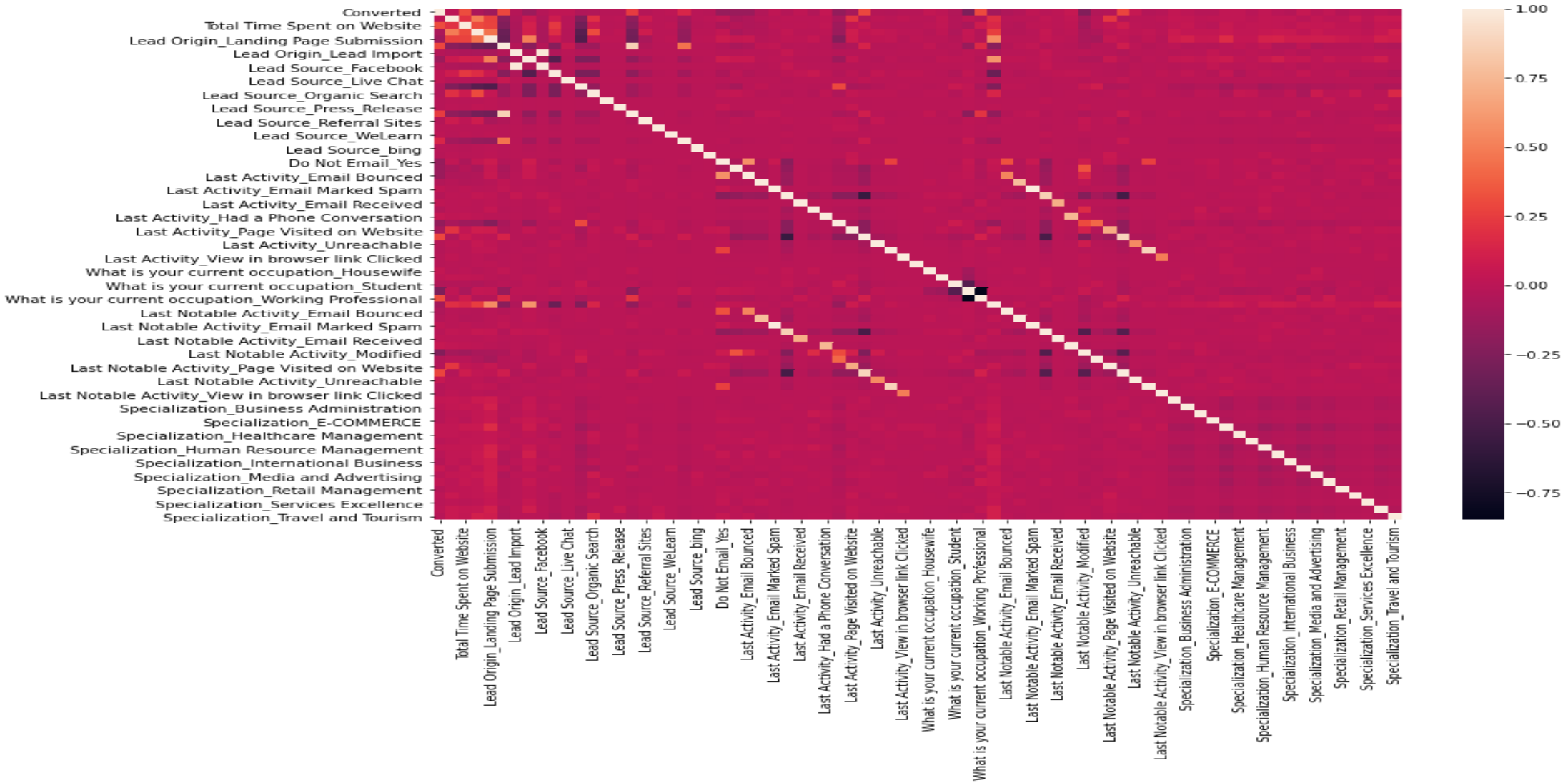


- Number of Unemployed leads are more than any other category.
- Working Professional is second highest among all.
- Others are lowest among all.

Pair Plots of leads



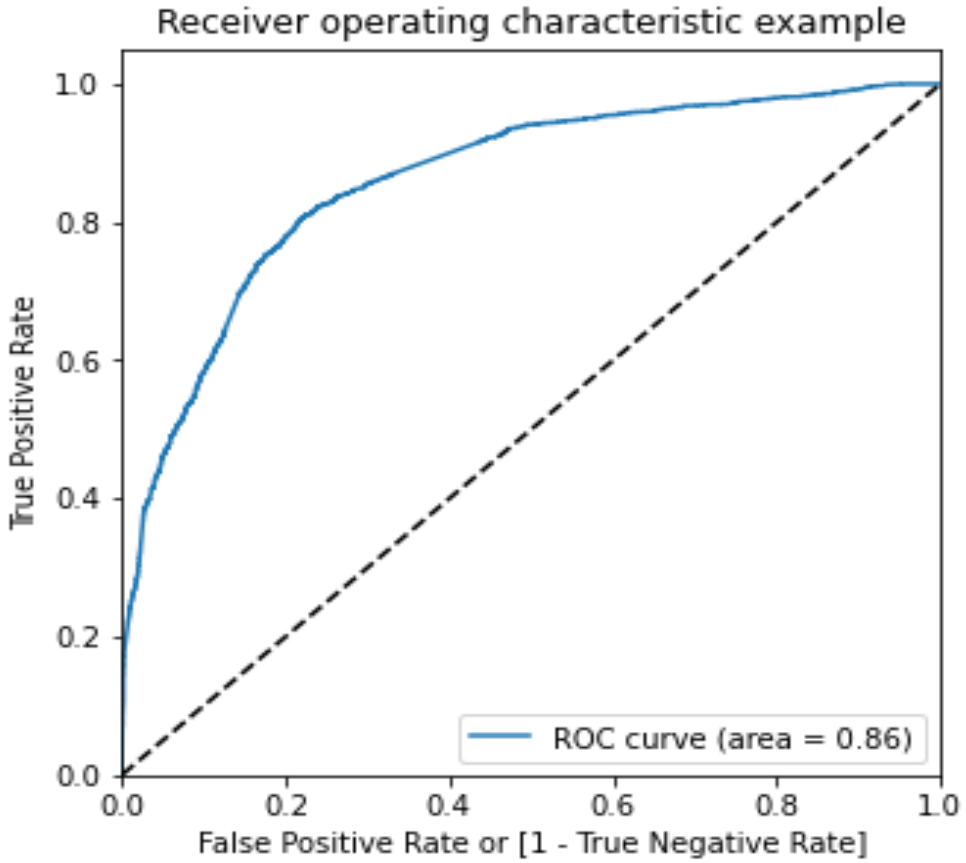
Heatmap of Leads



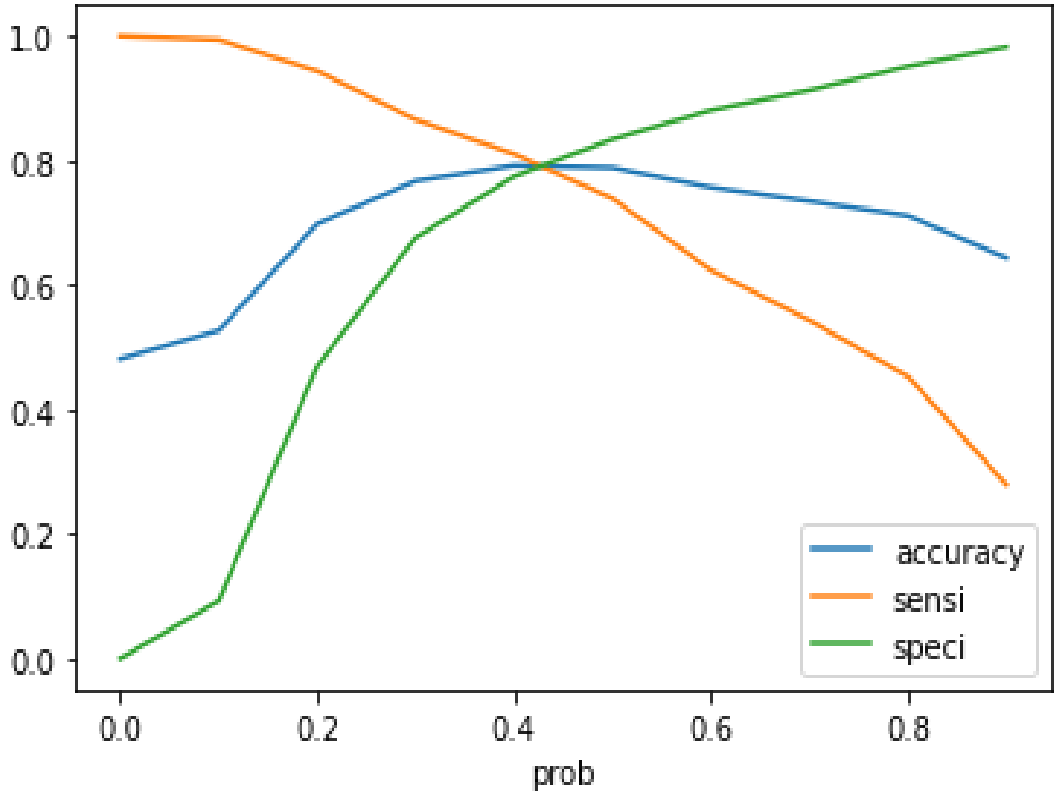
VIF data frame for all the variables present

	Features	VIF
9	What is your current occupation_Unemployed	2.82
1	Total Time Spent on Website	2.00
0	TotalVisits	1.54
7	Last Activity_SMS Sent	1.51
2	Lead Origin_Lead Add Form	1.45
3	Lead Source_Olark Chat	1.33
4	Lead Source_Welingak Website	1.30
5	Do Not Email_Yes	1.08
8	What is your current occupation_Student	1.06
6	Last Activity_Had a Phone Conversation	1.01
10	Last Notable Activity_Unreachable	1.01

both the p-values and VIFs seem decent enough for all the variables



The area under the curve of the ROC is 0.86 which is quite good. So we seem to have a good model. Let's also check the sensitivity and specificity tradeoff to find the optimal cutoff point.



The optimal values of the three metrics. So let's choose 0.42 as our cutoff now.

Summary

- The logistic regression model predicts the probability of the target variable having a certain value, rather than predicting the value of the target variable directly. Then a cutoff of the probability is used to obtain the predicted value of the target variable.
- Here, the logistic regression model is used to predict the probability of conversion of a customer.
- Any lead with greater than optimal value probability of converting is predicted as Hot Lead (customer will convert) and any lead with optimal value or less probability of converting is predicted as Cold Lead (customer will not convert).
- There are a lot of leads generated in the initial stage but only a few of them come out as paying customers from the bottom.
- In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.
- First, sort out the best prospects from the leads you have generated. 'Total Visits' , 'Total Time Spent on Website' , 'Page Views Per Visit' which contribute most towards the probability of a lead getting converted.
- Then, You must keep a list of leads handy so that you can inform them about new courses, services, job offers and future higher studies.
- The final model has Sensitivity of 0.78this means the model is able to predict 78% customers out of all the converted customers.
- The final model has Precision of 0.78 ,this means 78% of predicted hot leads are true leads.