Lead Score Case Study

OVERVIEW

COMPANY PROFILE

- An education company named X Education sells online courses to industry professionals.

BUSINESS CASE & OBJECTIVE

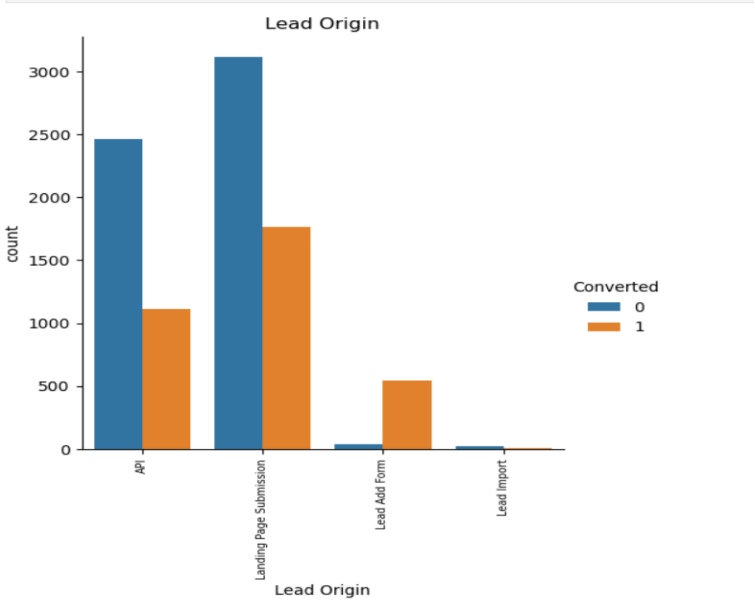
- Many professionals who are interested in the courses land on their website and browse for courses. They have process of form filling on their website after which the company that individual as a lead.
- Once these leads are acquired, employees from the sales team start making calls, writing emails, 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

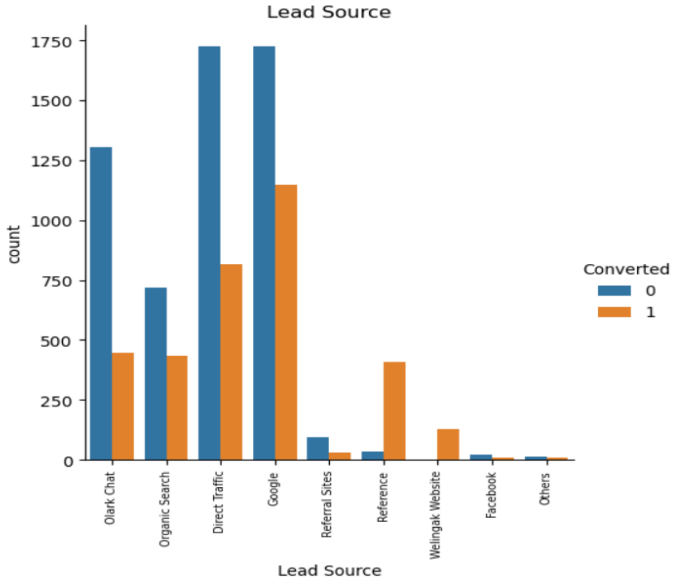
upGrad

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 pvalues)
- Model Evaluation
- Making predictions on test set

Exploratory Data Analysis

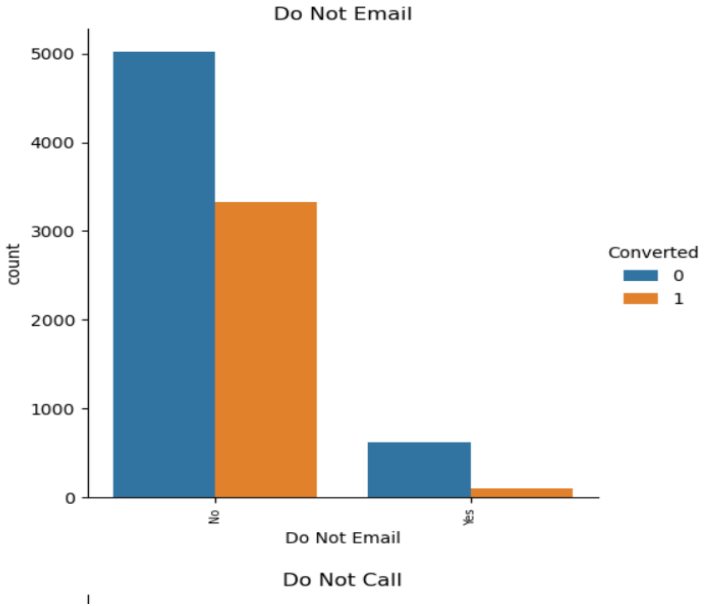




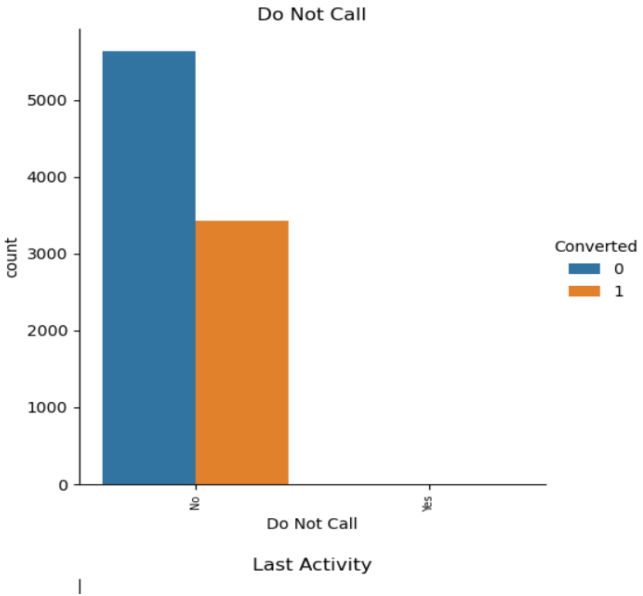
- Landing page submission has high lead conversions

- Google search has high conversions compared to other modes ,while references has high conversion rate.

Exploratory Data Analysis

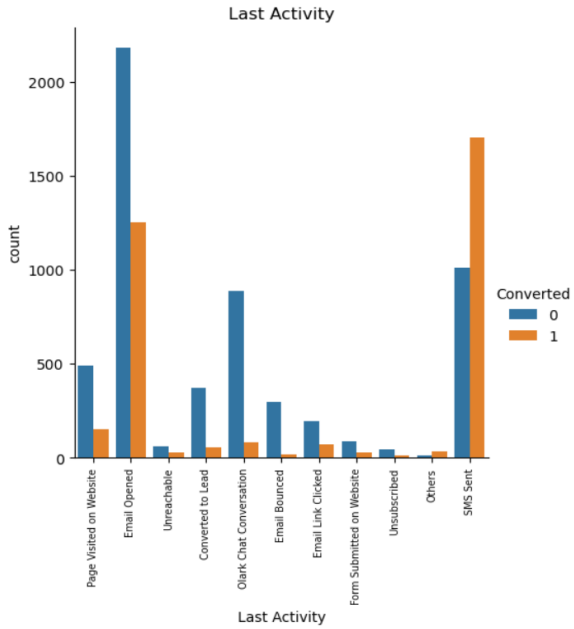


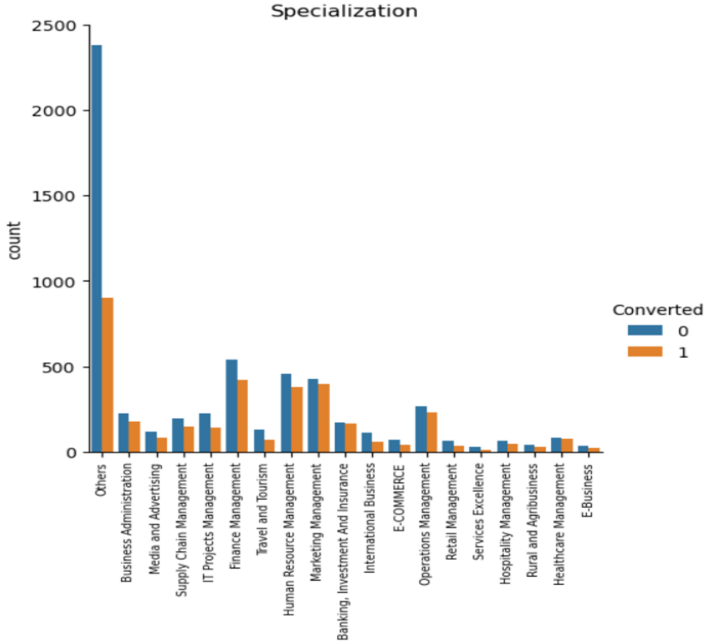
- Most of the leads prefer not to inform through Email



- Most of the leads prefer not to inform through Phone

Exploratory Data Analysis

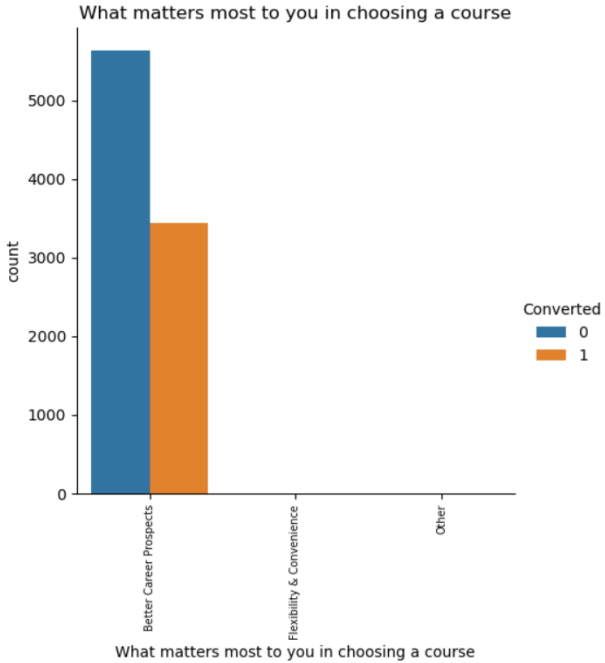


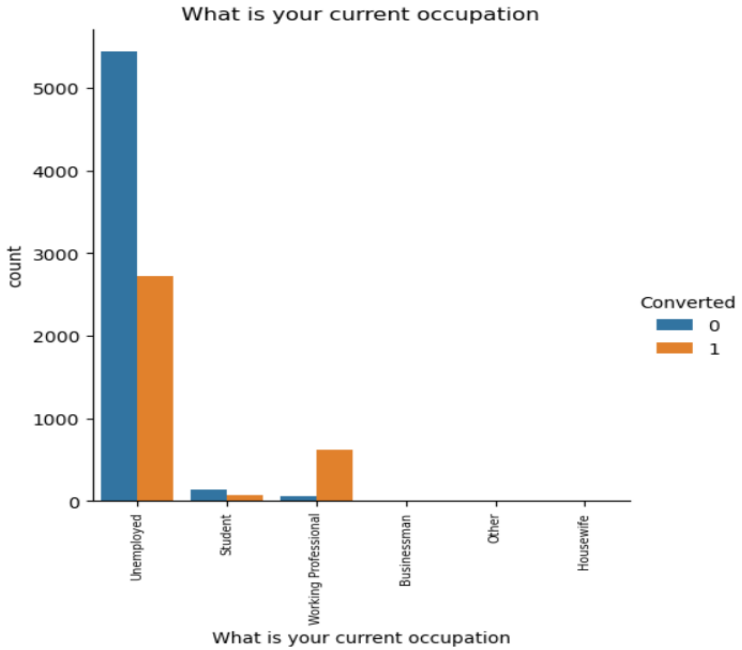


- SMS has shown to be a promising method for getting higher method lead.

- Most of the leads have no informations(others) about specialization.

Exploratory Data Analysis

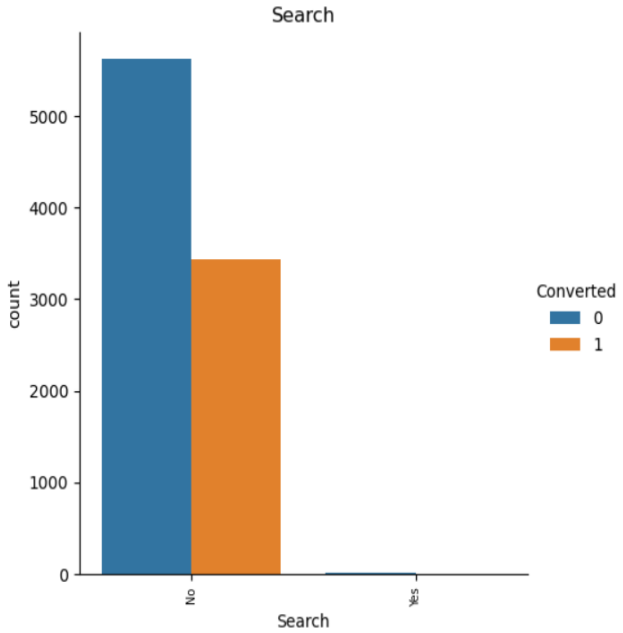




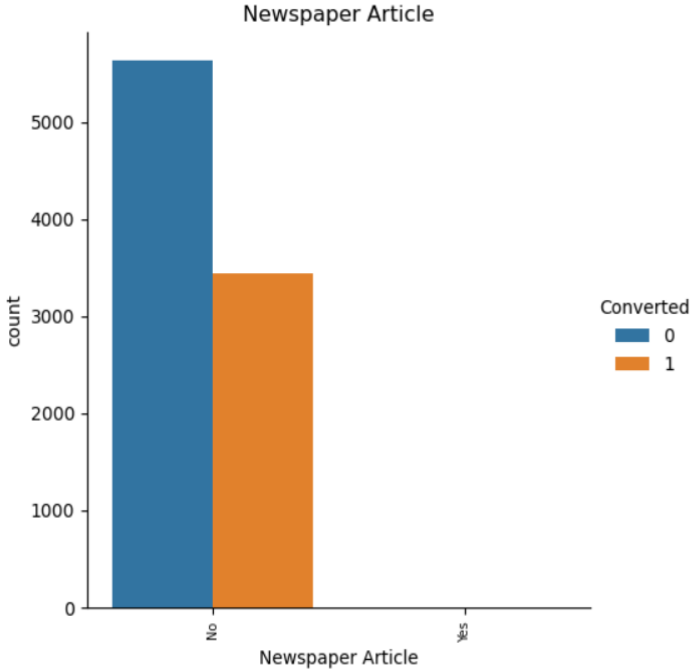
- Most of the leads are looking for better career opportunity

- Leads who are unemployed are opting for course.

Exploratory Data Analysis

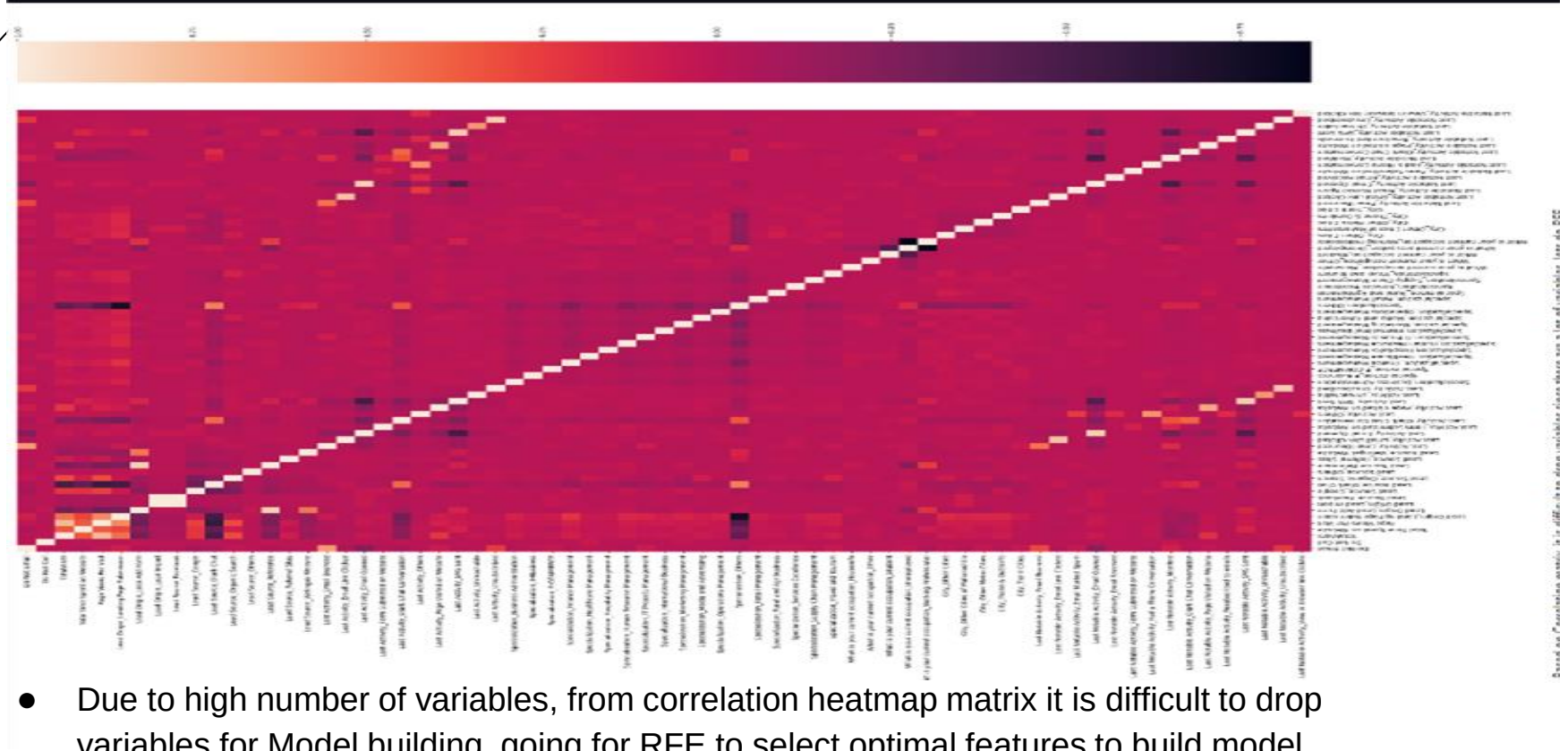


- The above graph shows searches are not good source of lead



- The above graph shows Newspaper article are not good source of lead

Correlation



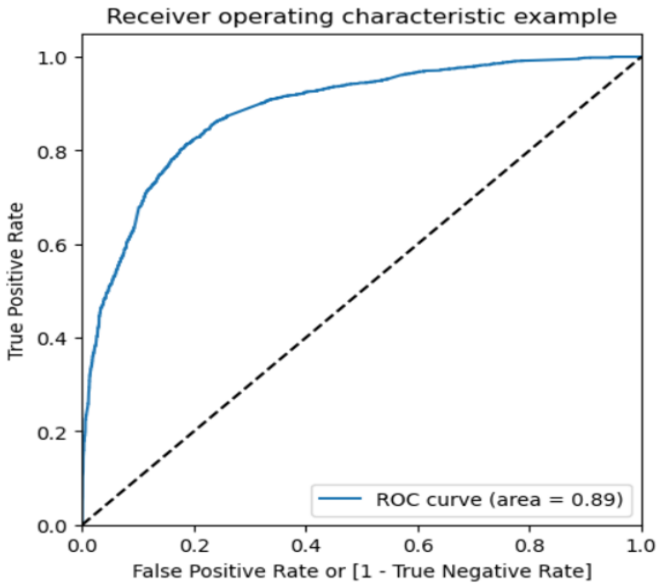
- Due to high number of variables, from correlation heatmap matrix it is difficult to drop variables for Model building, going for RFE to select optimal features to build model

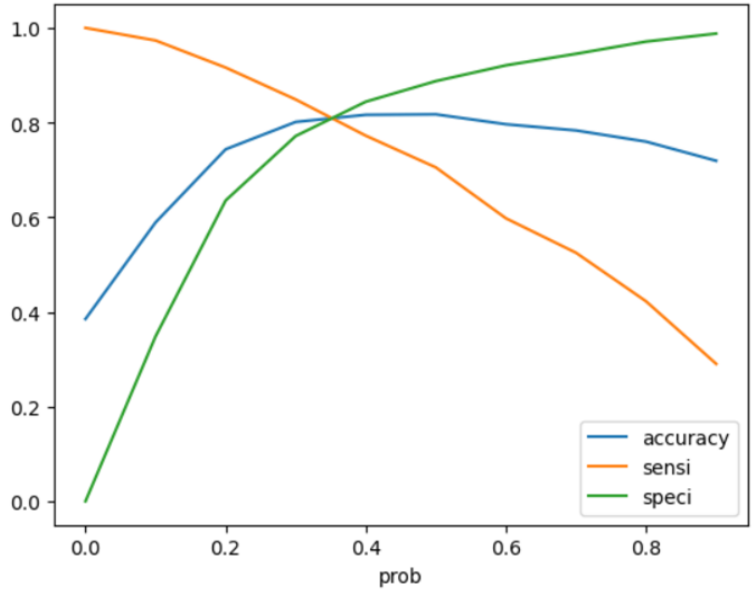
Model Building

- Splitting the Data into Training and Testing Sets
- Scale variable in train set
- Build the first model
- Use RFE to eliminate less relevant variable
- Build next model
- Predict using train set
- Evaluate accuracy and other metrics
- Predict using test set
- Precision and recall analysis on prediction test

- Model 9 is our final model with 12 variables, with accuracy and specificity around 80% and sentivity around 70%. An arbitary cut off of 0.5 is taken and these 3 metrics are arrived at.

ROC curve





- ROC gave us us (0.89) AUC (area under curve), therefore our model is a pretty good one.
- From the curve above, 0.34 is the optimum point to take it as a cutoff probability and sentivity improved to ~82%.

Observation

After running the model on the Test Data , we obtain:

- Accuracy : 80.2 %
- Sensitivity : 79.9 %
- Specificity : 80.4 %

Results : Comparing the values obtained for Train & Test:

** Train Data:

- Accuracy : 81.1 %
- Sensitivity : 81.7%
- Specificity : 80.7 %

** Test Data:

- Accuracy : 80.2 %
- Sensitivity : 79.9 %
- Specificity : 80.4 %

Conclusion

EDA:

- SMS message have higher impact on conversions.
- Landing Page submission can help find out more leads
- Highest number of leads are generated by Google/direct traffic.

Logistic regression Model:

- The model shows high close to 80.2% of accuracy
- The model shows 81.7% accuracy and 80.4% of specificity.
- Overall this model proves to be accurate.

THANK YOU

upGrad