Lead Scoring Case Study

Group Members:

1. Manish Kumar Dabhade
2. Dharan R
3. Deepak Raja S

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.

Objectives

- X education wants to select most promising leads, i.e. the leads that are most likely to convert into paying customers.
- Company wants to build a model so as to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.
- Deployment of the model for the future use.

Adopted Methodology

- Data cleaning and data manipulation.
- EDA
- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

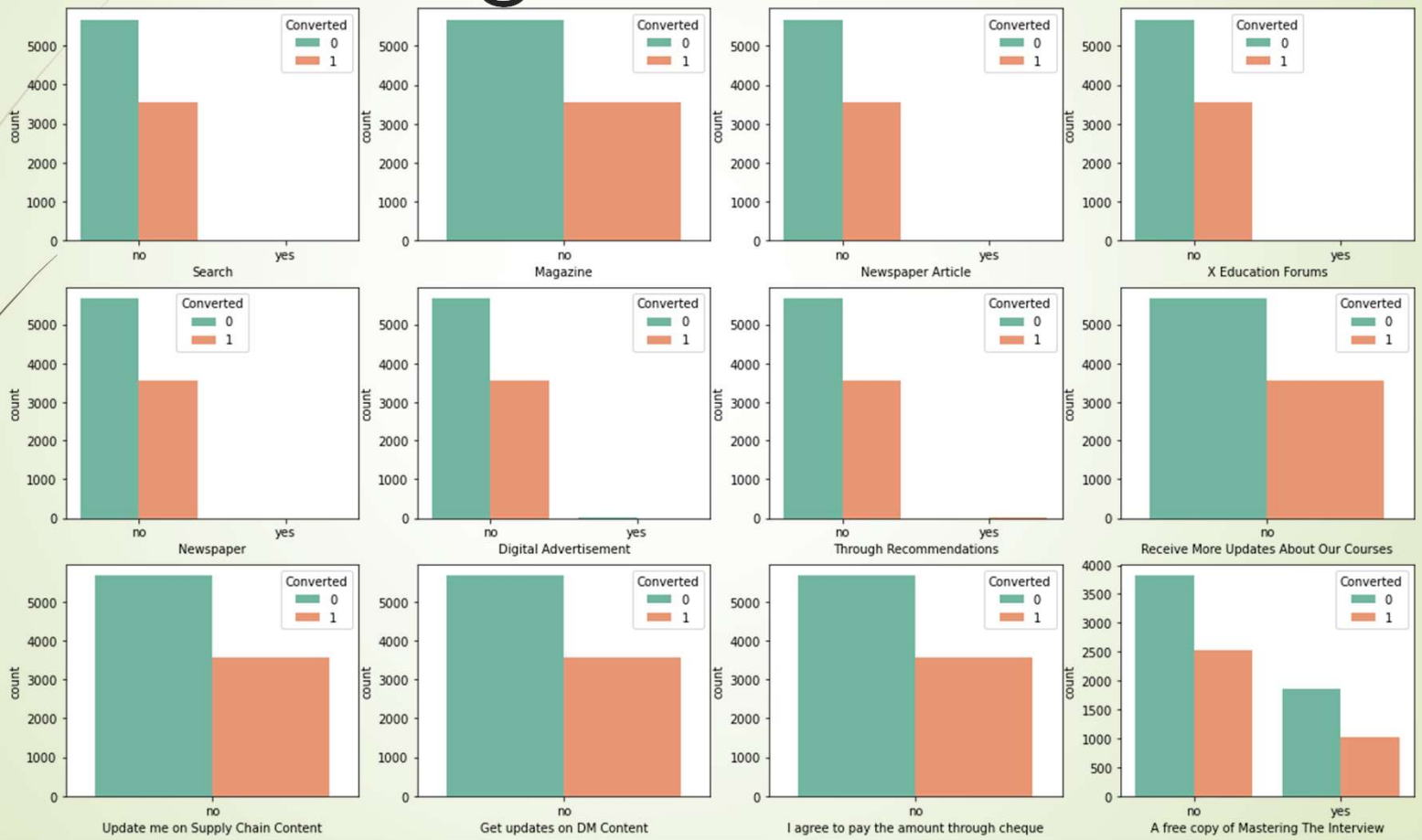
Data cleaning and Data manipulation

- Converting all the values to lower case.
- Dropping ‘Lead Number’ and ‘Prospect ID’ as they have all unique values.
- Replacing 'Select' values with NaN.
- Handling Missing Values.

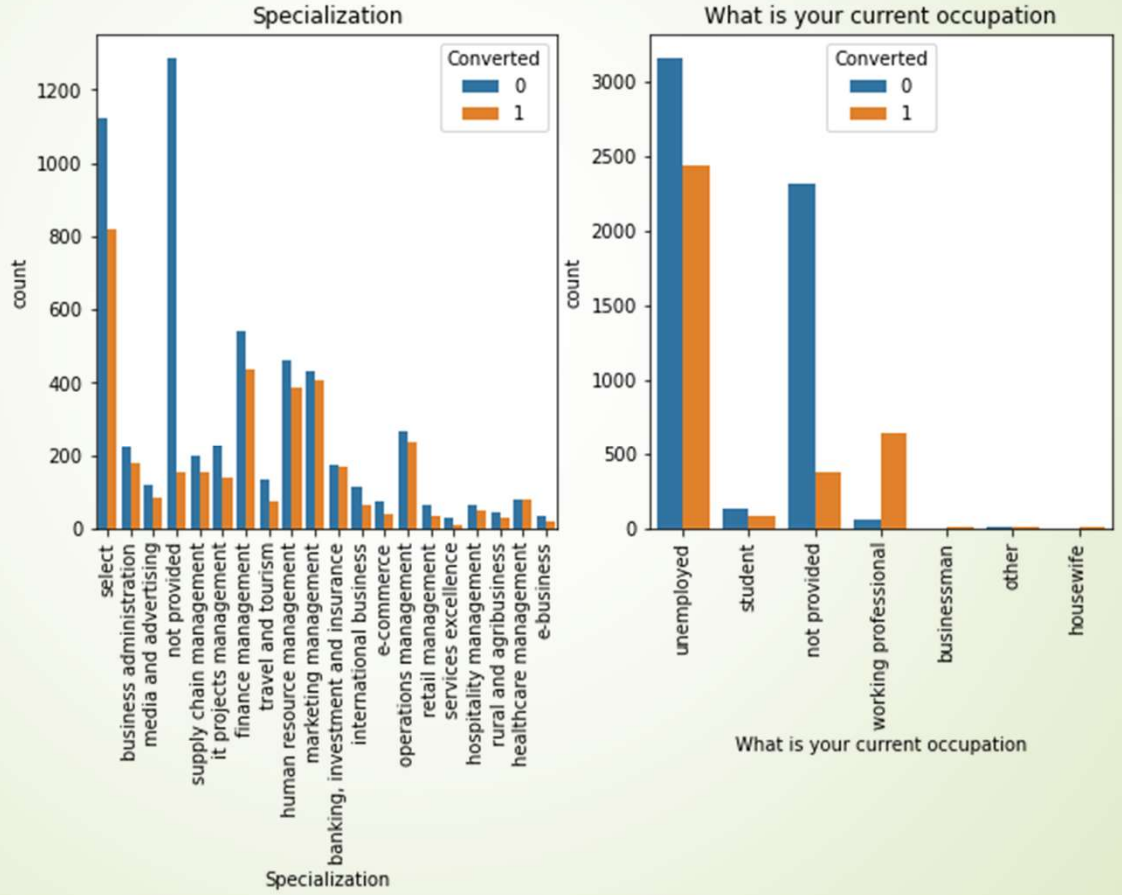
Exploratory Data Analysis

- Performed categorical attribute analysis.
- Imbalance in the data was not observed.
- Checked the correlation among variables using heatmap.
- Creating dummy variables.

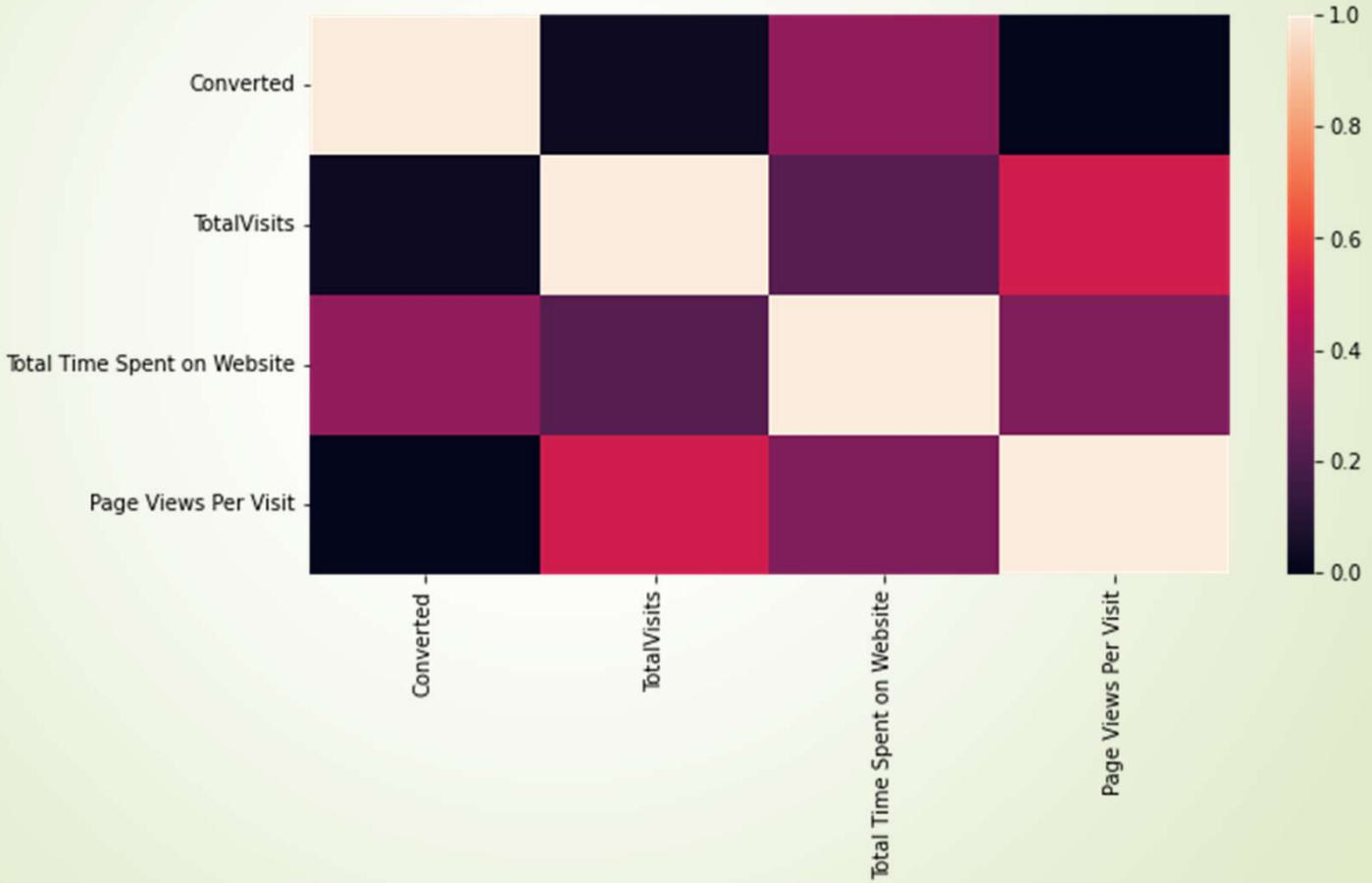
Visualzing variables for imbalancing



Plot of Count Vs Specialization and ‘What is your current occupation



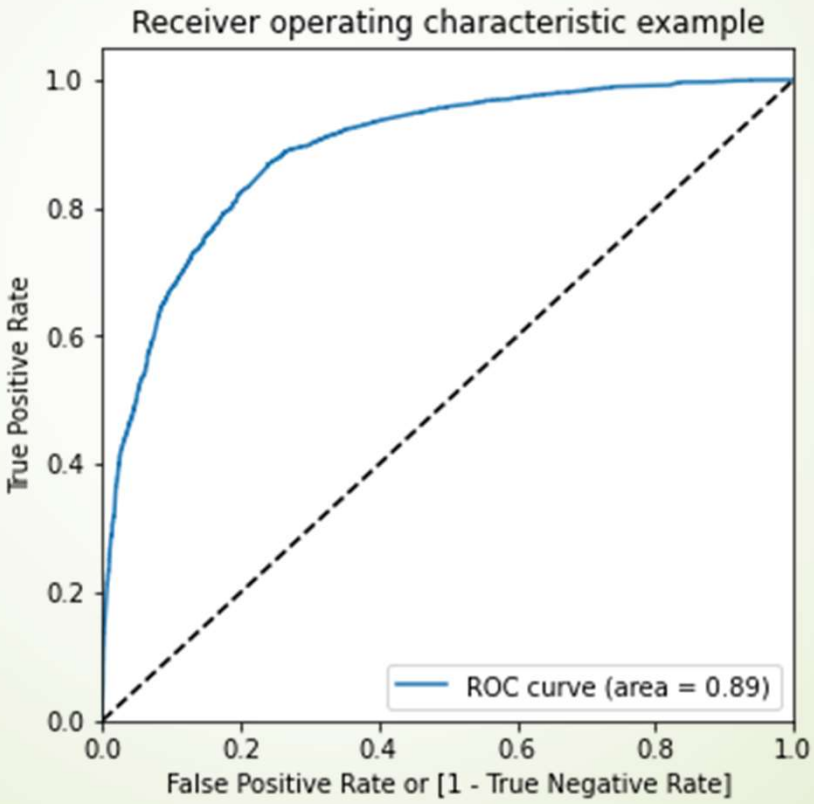
Correlation among variables



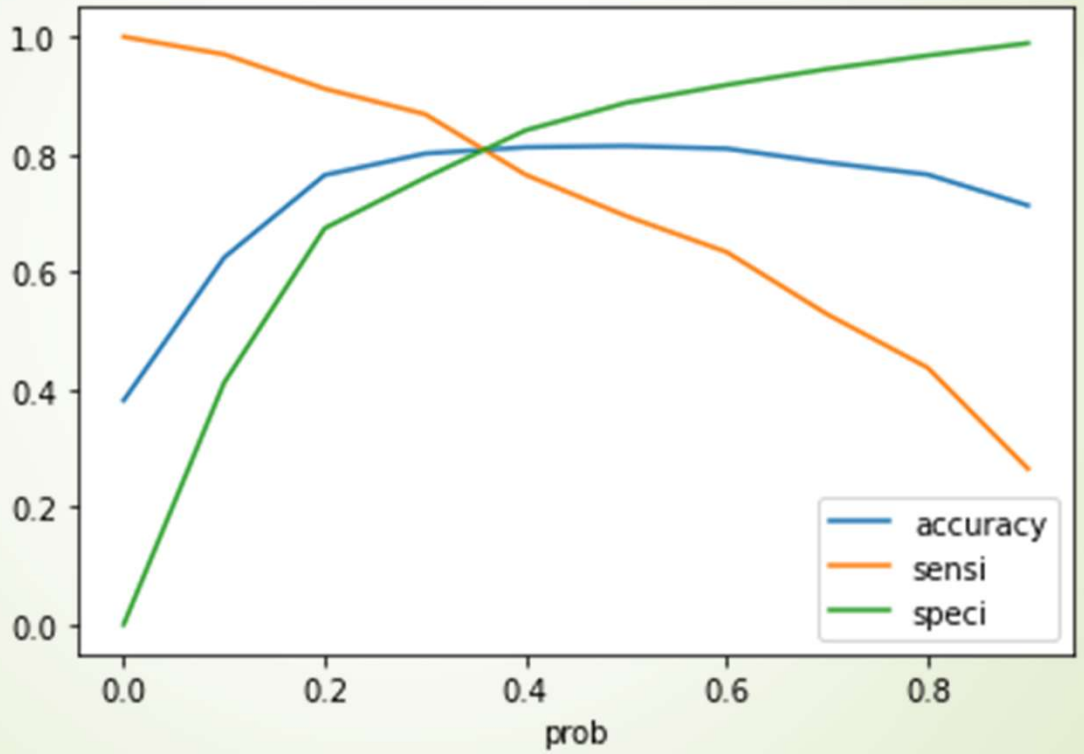
Model Building

- Dividing the data into train and test data sets in the ratio of 70% and 30%.
- Performing feature scaling.
- Use RFE for feature selection.
- Running RFE with 15 variables as output.
- Moving forward with stable model 4 having significant p-values.
- Calculation VIF values.
- Creating predictions and metric calculation.
- Optimizing Cut-off (ROC Curve).
- Performing predictions on Test set.

ROC Curve


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.89)
0.0
0.2
0.4
0.6
0.8
1.0

Plot of Accuracy, Sensitivity and Specificity



Observations

- It was found that the variables that mattered the most in the potential buyers
  - (a) Total time spent on website.
  - (b) Total number of visits.
- Current occupation is as a working professional, Direct traffic, Google, 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.

Conclusion

- Accuracy, Sensitivity and Specificity values of test set are around 79%, 91% and 71% which are approximately closer to the respective values calculated using trained set.
- We have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Hence overall this model seems to be good.