LEAD SCORING CASE STUDY

Submitted by
1 Sheetal pandey
2 Shikhar Rai
3 Shashwat Krishna

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.

Business Objective:

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

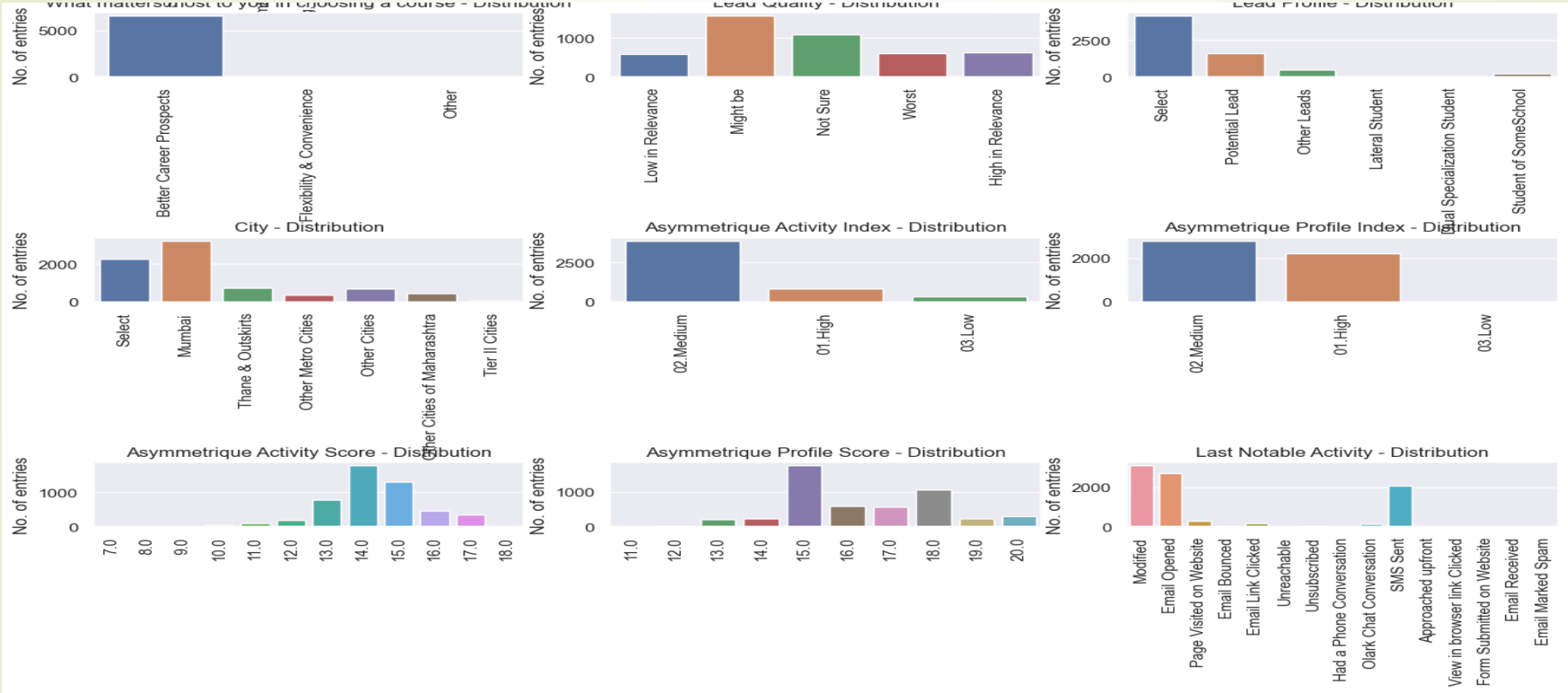
Solution Methodology

- Data cleaning and data manipulation.
  1. Check and handle duplicate data.
  2. Check and handle NA values and missing values.
  3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
  4. Imputation of the values, if necessary.
  5. Check and handle outliers in data.
- EDA
  1. Univariate data analysis: value count, distribution of variable etc.
  2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.
- 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 Manipulation

- Total Number of Rows =37, Total Number of Columns =9240.
- Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply”
- Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.
- Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.

Exploratory DATA ANALYSIS

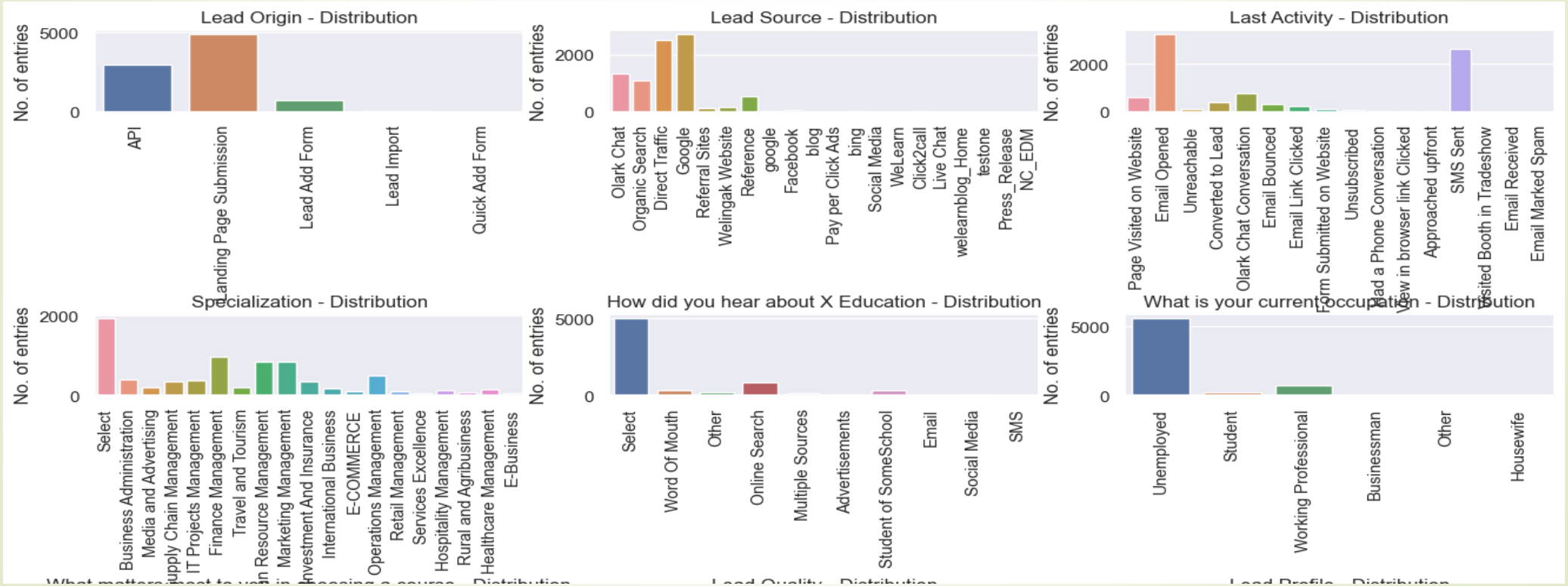


-There are a few columns in which there is a level called 'Select' which is taking care

-Leads from HR, Finance & Marketing management specializations are high probability to convert.

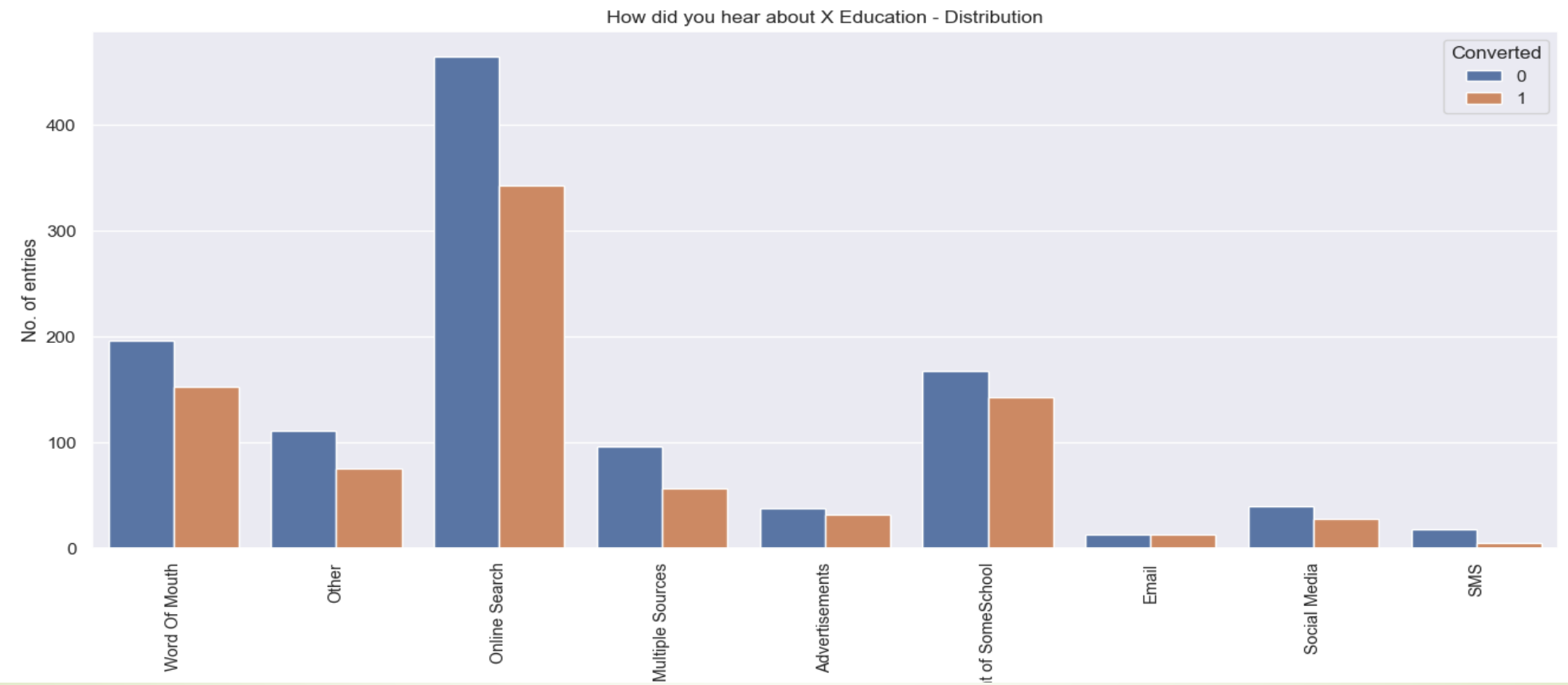
-In lead source the leads through google & direct traffic high probability to convert

-Leads which are opening email have high probability to convert, Same as Sending SMS will also benefit.



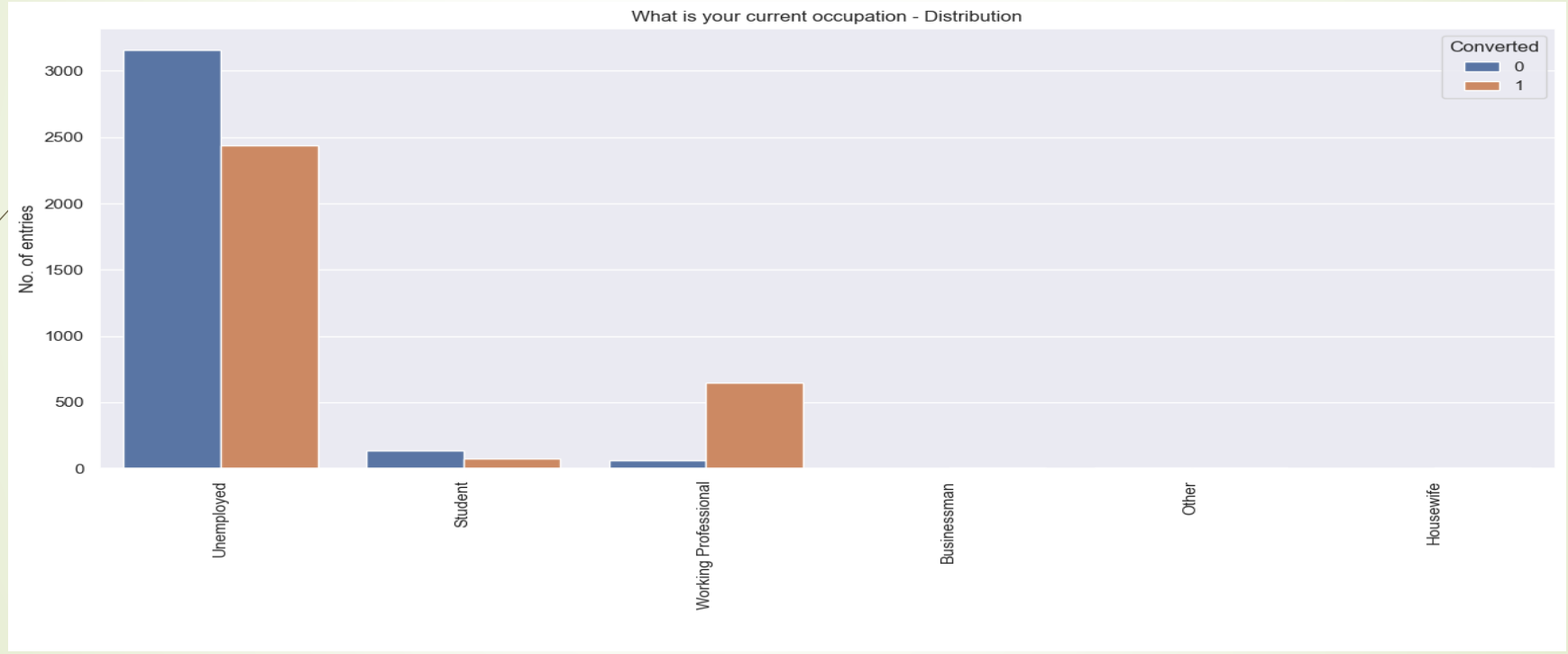
How did you hear about X education

Leads Mostly come from online sources

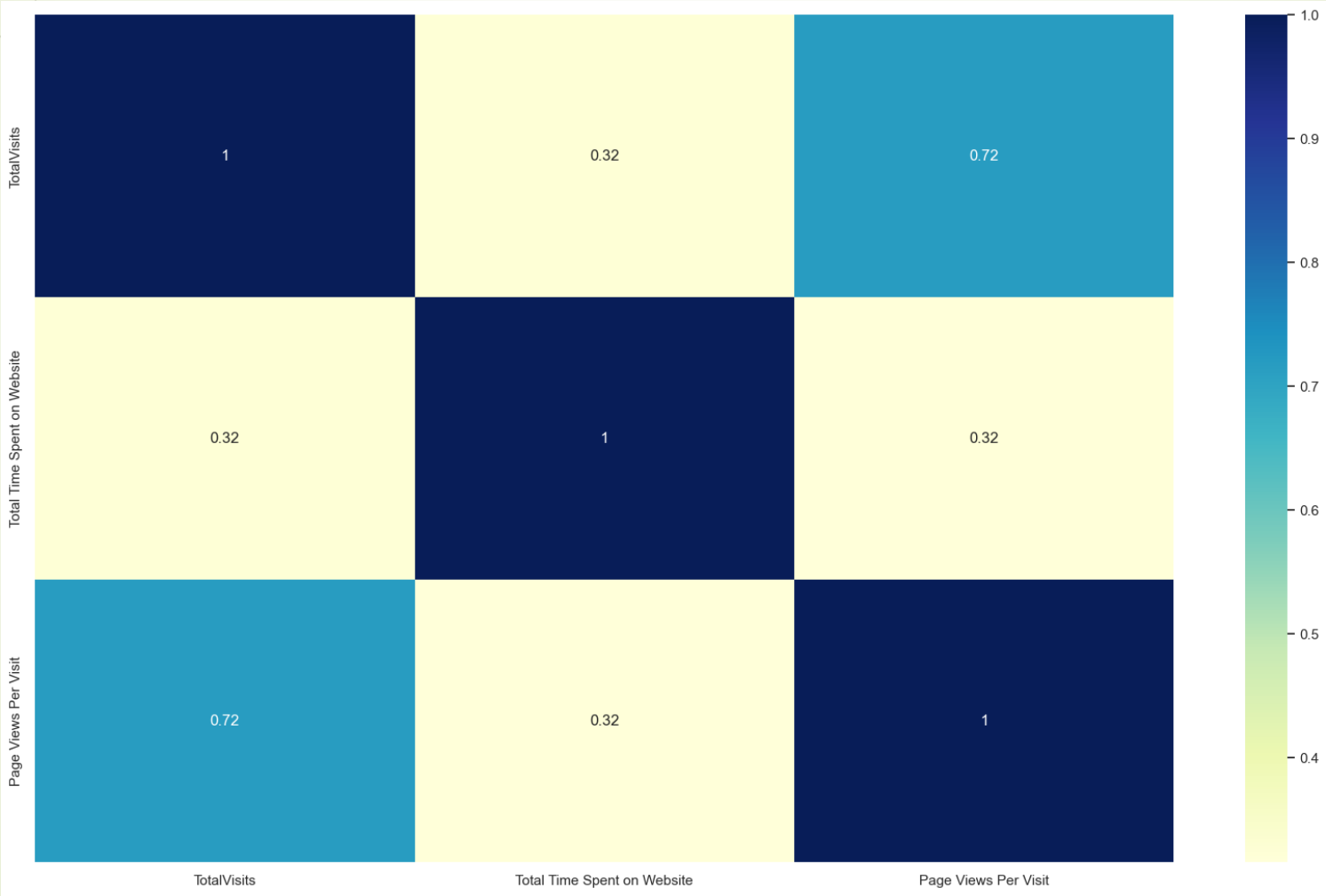


What is Your Occupation

Leads which are Unemployed are more interested to join the course than others.



Correlation



Data Conversion

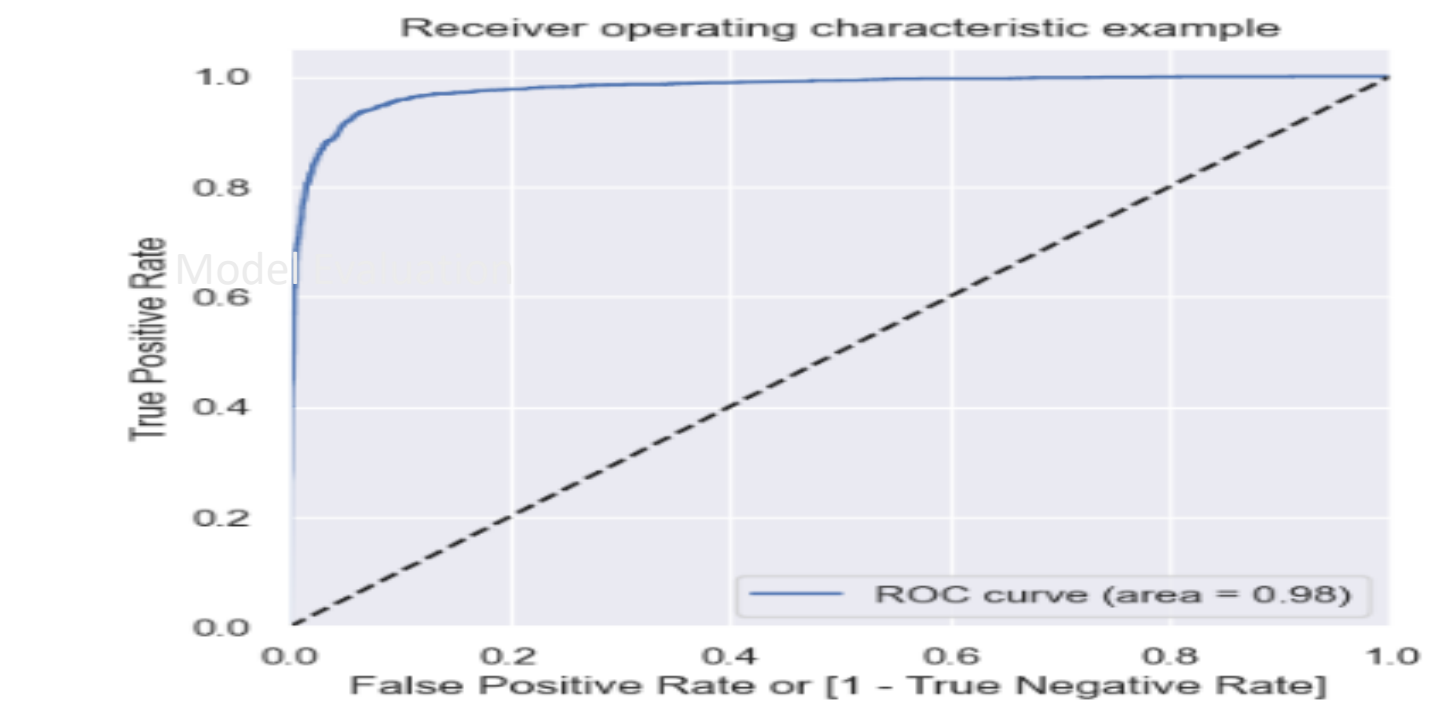
- Numerical Variables are Normalised
- A dummy variable is one that takes a binary value to indicate the absence or presence of some categorical effect that may be expected to shift the outcome.

Model Building

- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
  - Use RFE for Feature Selection
  - Running RFE with 15 variables as output
  - Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5
  - Predictions on test data set

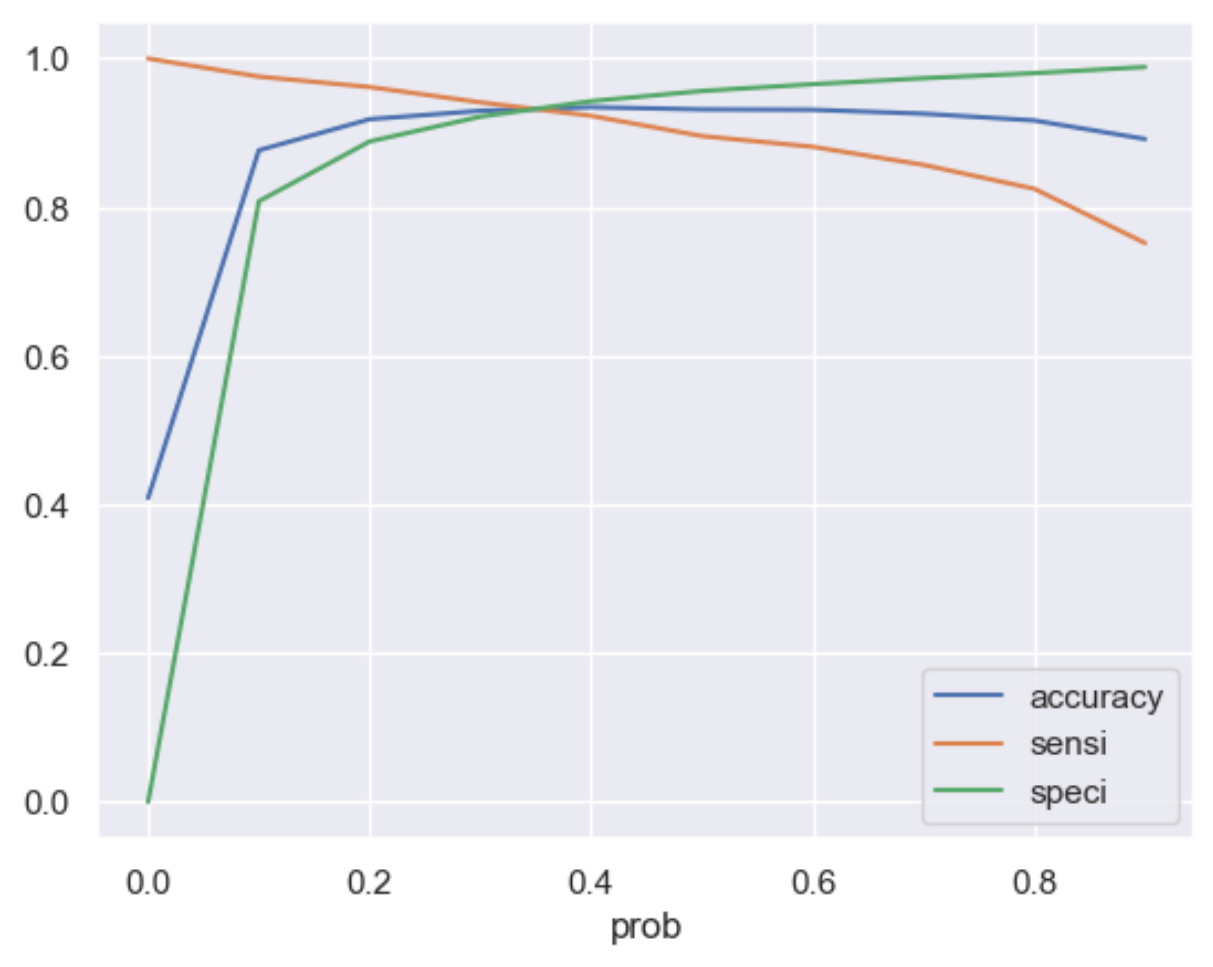
Model Evaluation

ROC Curve



- **The MODEL has Area Under Curve (AUC) value of 0.98, which is a very good indicator**

The various cutoff are plotted and 0.38(approx 0.4) is where all values are converging.



Conclusion

Area Under Curve (AUC) value of 0.98

Scores on Train set
- Sensitivity: 92.29 %
- Specificity: 94.27 %
- ACCURACY SCORE: 93.46 %
- Precision Score of the Model: 91.76 %
- Recall Score of the Model: 92.29 %

Scores on Test set
- Sensitivity of the Test Predictions: 91.86 %
- Specificity of the Test Predictions: 93.75 %
- Precision Score of the Test Data Predictions: 90.6 %
- Recall Score of the Test Data Predictions: 91.86 %
- Accuracy of the Test Data Predictions: 93.0 %

Final score

Positive Predictors

- A customer with these TAGS assigned is a potential Lead: "Closed by Horizzon", "Lost to EINS", "Will revert after reading the email"
- A customer Lead sourced by "Welingak Website" is a Hot Lead.
- A customer who is currently "Working Professional" or "Unemployed" is a Hot Lead.

•Lead Profile_Lateral Student- 25.4200
•Tags_Closed by Horizzon: 6.0962
•Lead Source_Welingak Website: 5.1297
•Tags_Lost to EINS: 4.9405
•Tags_Will revert after reading the email: 3.1744
•What is your current occupation_Working Professional: 1.98
•What is your current occupation_Unemployed: 1.5563

Negative Predictors

- A customer with these TAGS assigned is NOT a potential Lead: "Already a Student", "switched off", "Not doing further education", "Diploma holder (Not Eligible)", "Ringing", "Interested in other courses", "Interested in full time MBA"
- A customer whose Lead Quality is deemed as "Worst" is also NOT a Hot Lead.

•const: -3.9500
•Tags_Already a student: -4.0948
•Tags_switched off: -4.4765
•Tags_Not doing further education: -3.0528
•Lead Quality_Worst: -3.6051
•Tags_Diploma holder (Not Eligible): -2.9483
•Tags_Ringing: -4.1671
•Tags_Interested in other courses: -2.7307
•Tags_Interested in full time MBA: -2.5468

Thank You