LEAD SCORE CASE STUDY

Presented by:
Manali Pote

Agenda

PROBLEM STATEMENT

SOLUTION METHODOLOGY

DATA MANIPUTALTION

MODEL BUILDING

CONCLUSION

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 leads rather than making calls to everyone.

BUSINESS GOAL:

- X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company needs a model in which each lead is given a score, with the goal being to increase the conversion chance for consumers with higher lead scores and decrease it for those with lower lead scores.
- Specifically, the CEO has stated that an approximate 80% lead conversion rate is the goal.

SOLUTION APPROACH

❑DATA CLEANING AND MANIPULATION:

1. Check and handle duplicate data.
2. Check and handle NA values and missing values.
3. Drop columns, if it contains large number of missing values and not useful for the analysis.
4. Imputation of the values, if necessary.
5. Check and handle outliers in data.

❑EDA:

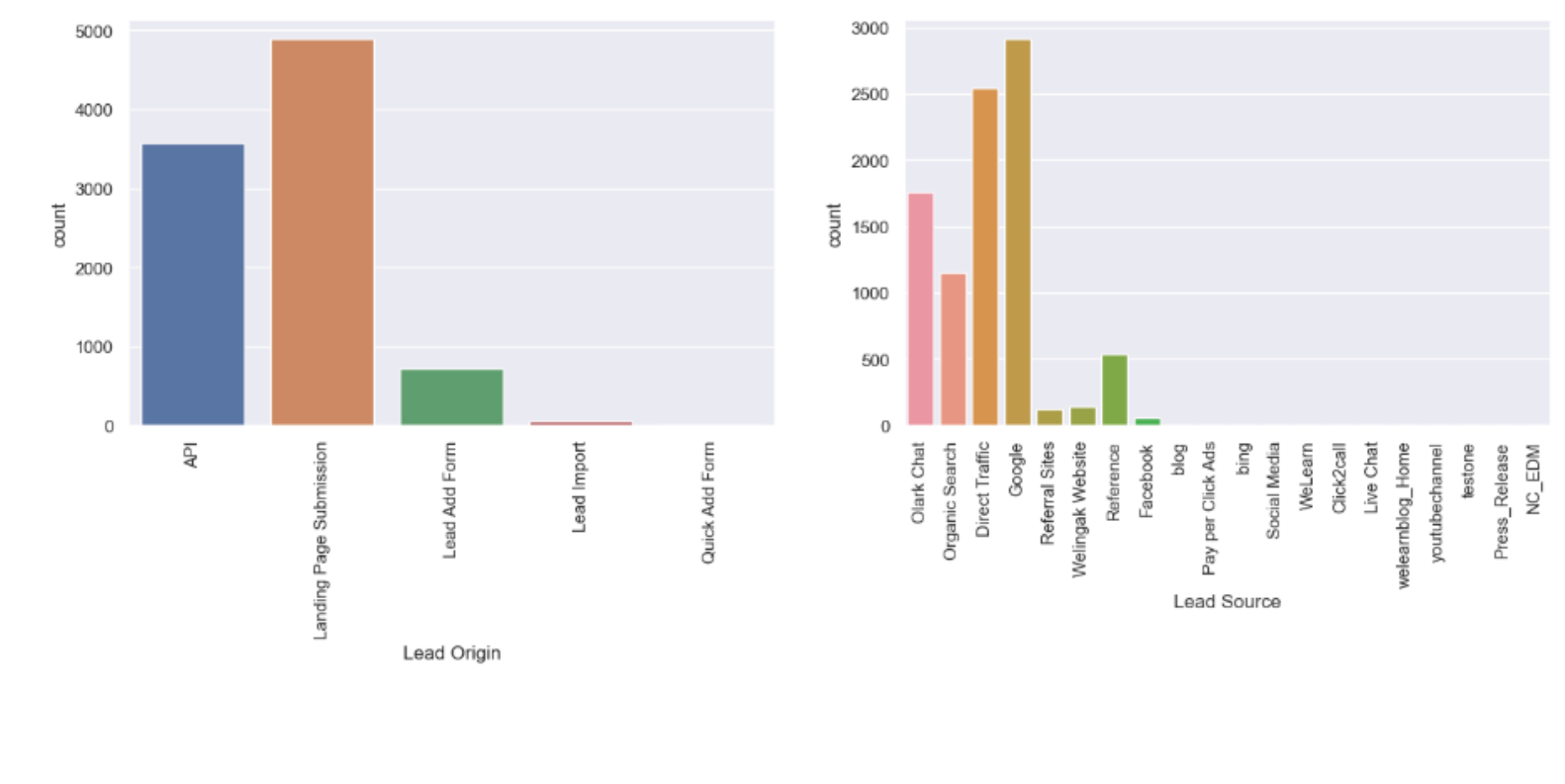
1. Univariate data analysis: value count, distribution of variable etc.
2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.

- 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.

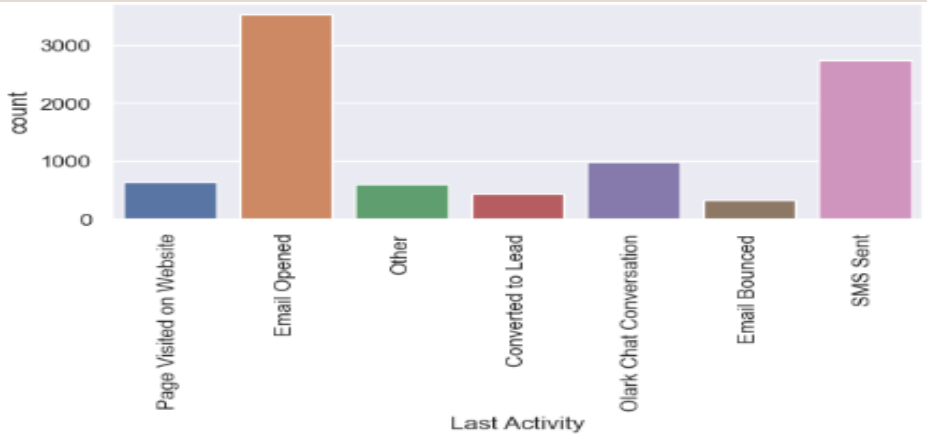
DATA MANIPULATION

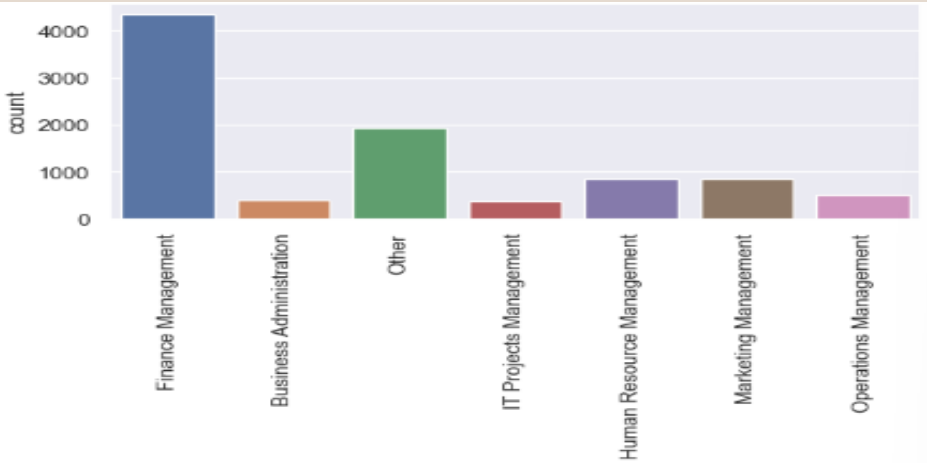
- Total Number of Rows =37, Total Number of Columns =9240.
- Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply”, “Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc.
- Replacing the select values with nan values.
- Checking percentage of null values in each columns.
- Dropping the columns having more than 40% of missing value such as ‘How did you hear about X Education’, ‘Lead Quality’, ‘Lead Profile’, ‘Asymmetrique Activity Index’, ‘Asymmetrique Profile Index ‘, ‘Asymmetrique Activity Score’, ‘Asymmetrique Profile Score ‘, ect.

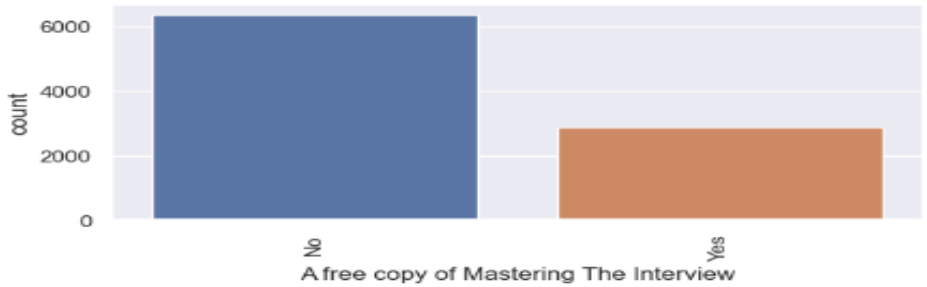
EDA

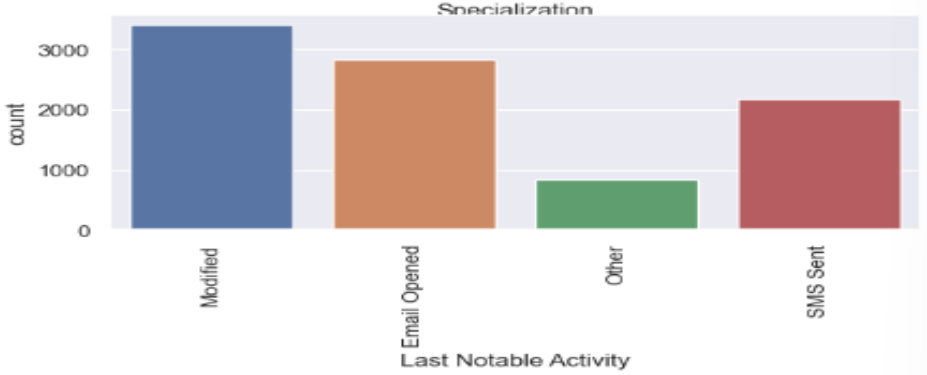

count
5000
4000
3000
2000
1000
0
API
Landing Page Submission
Lead Add Form
Lead Import
Quick Add Form
Lead Origin
count
3000
2500
2000
1500
1000
500
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Welingak Website
Reference
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM
Lead Source

BIVARIATE ANALYSIS

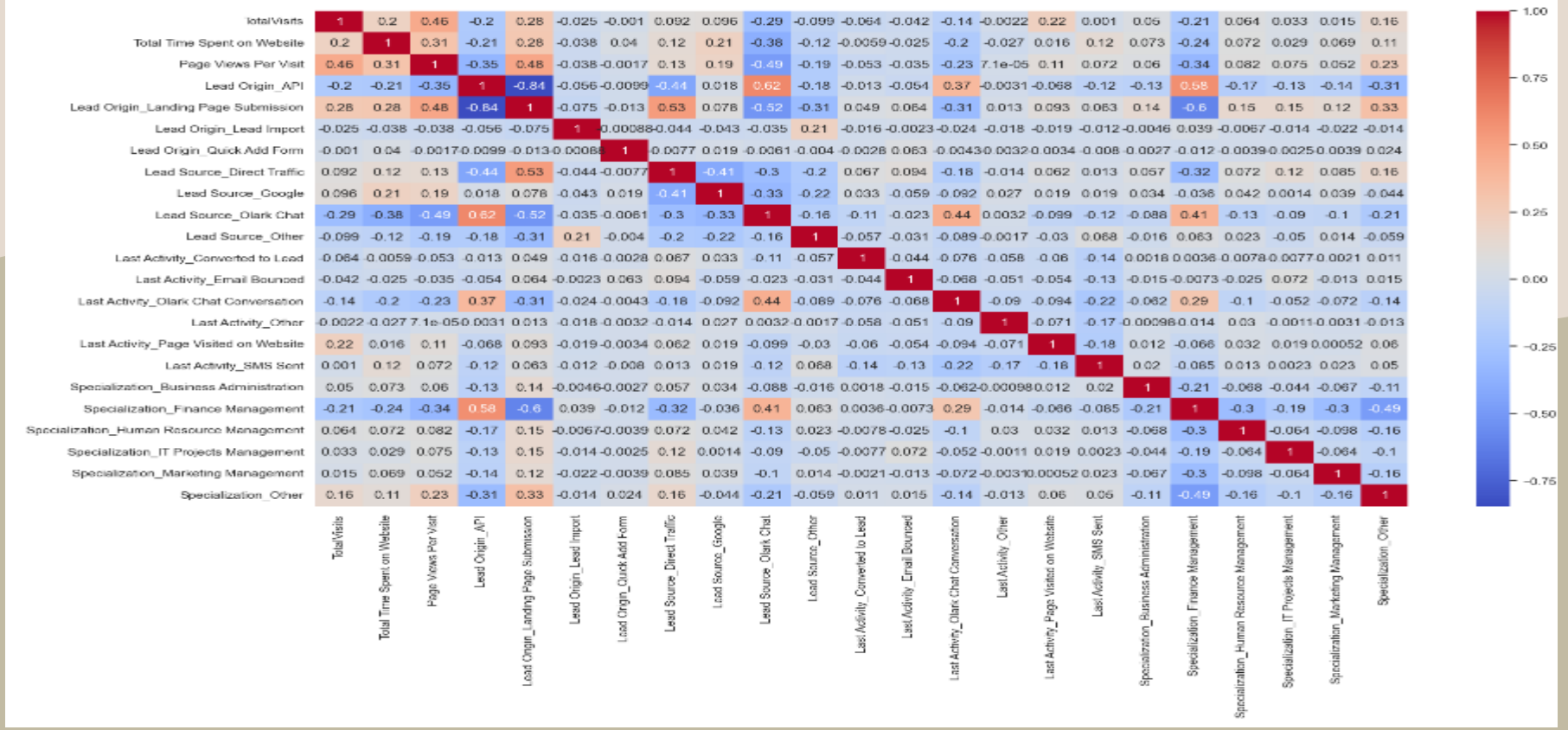








CORRELATION



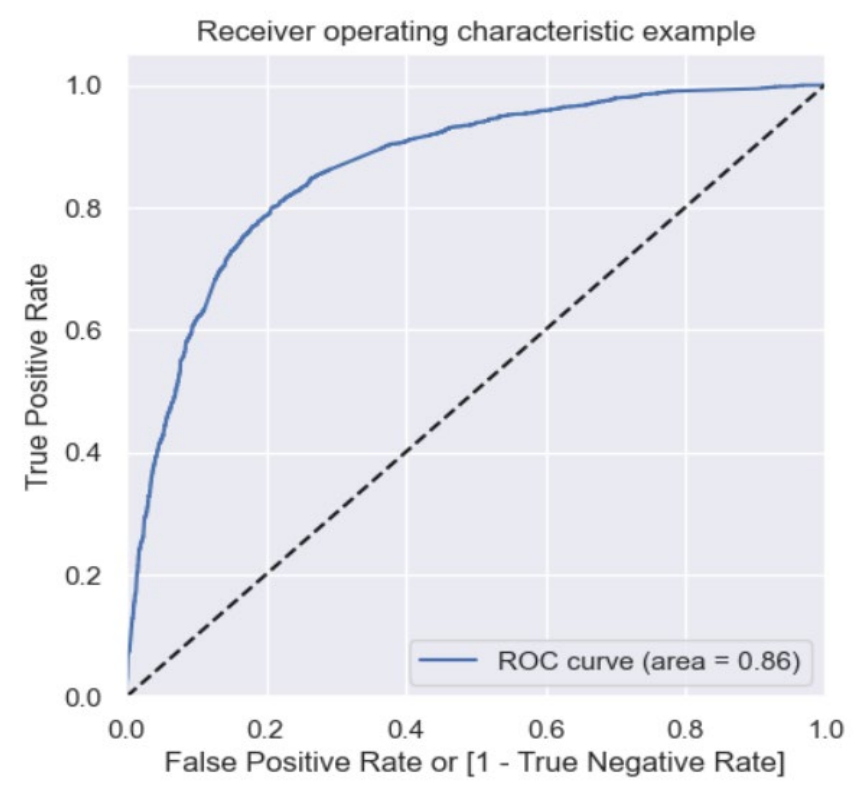
DATA CONVERSION

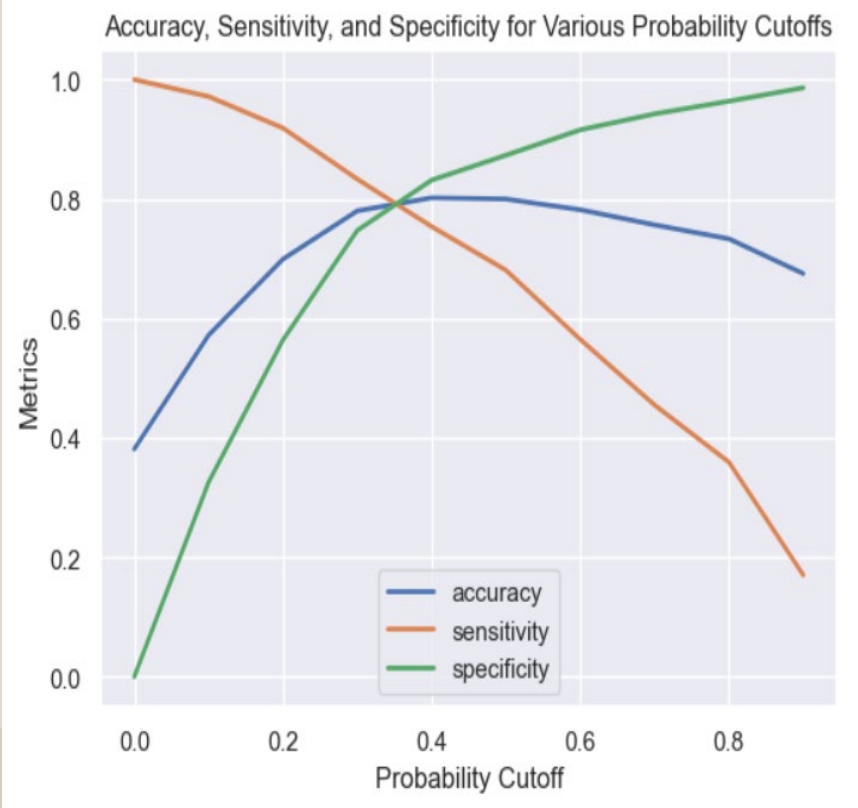
- Numerical variables are normalised.
- Dummy variables are created for object type variables.
- Total rows for analysis are 9240
- Total columns for analysis are 37

MODEL BUILDING

- Splitting the Data into Training and Testing Sets.
- We have selected a 70:30 ratio for our train-test split, which is the first fundamental stage in regression.
- Use RFE for feature selection.
- 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.
- Overall accuracy 80%

ROC CURVE


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.86)
0.0
0.2
0.4
0.6
0.8
1.0


Accuracy, Sensitivity, and Specificity for Various Probability Cutoffs
Metrics
Probability Cutoff
accuracy
sensitivity
specificity
0.0
0.2
0.4
0.6
0.8
1.0

- Finding optimal cut off point.
- Optimal cut off probability is that probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.37.

CONCLUSION

It was found that the variables that mattered the most in the potential buyers are (In descending order) :

- The total time spend on the Website.
- Total number of visits.
- When the lead source was:
  a. Google
  b. Direct traffic
  c. Organic search
  d. Welingak website
- When the last activity was:
  a. SMS
  b. Olark chat conversation

- When the lead origin is Lead add format.
- When their current occupation is as a working professional.
- 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.

Thank you