Lead Scoring Case Study

CONTENTS

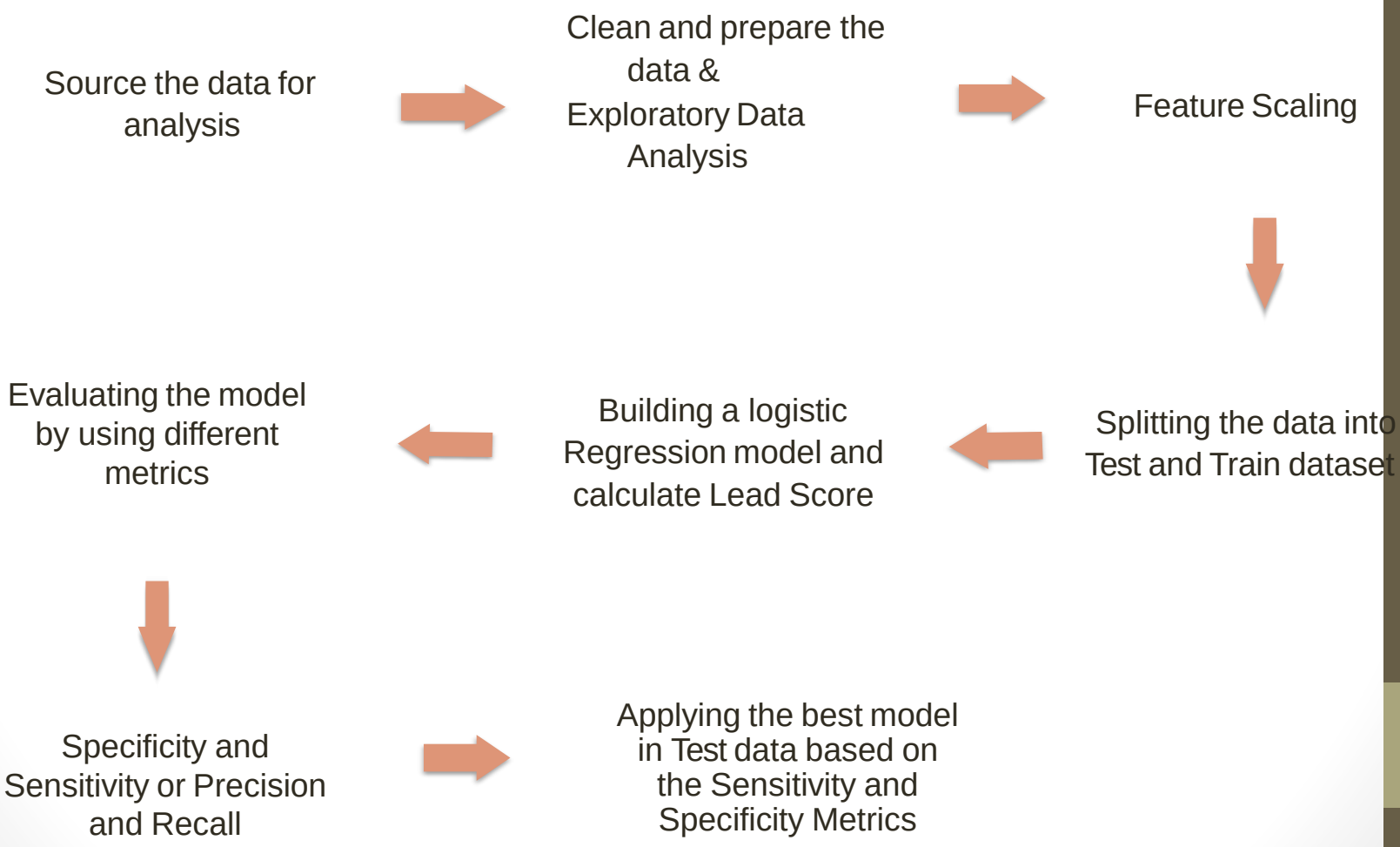
PROBLEM STATEMENT

- X Education sells online courses to industry professionals, gets a lot of leads but its lead conversion rate is very poor.
- To make this process more efficient, the company wishes to identify the leads with the most potential, also known as **'Hot Leads'**.
- If they successfully identify this set of leads, the lead conversion rate should spike since they would be communicating with the potential leads rather than everyone (which is troublesome and unfruitful)

BUSINESS OBJECTIVE

- To help X Education to select the most promising leads(Hot Leads), i.e. the leads that are most likely to convert into paying customers.
- To build a logistic regression model to assign a lead score value between 0 and 100 to each of the leads which can be used by the company to target potential leads

METHODOLOGY



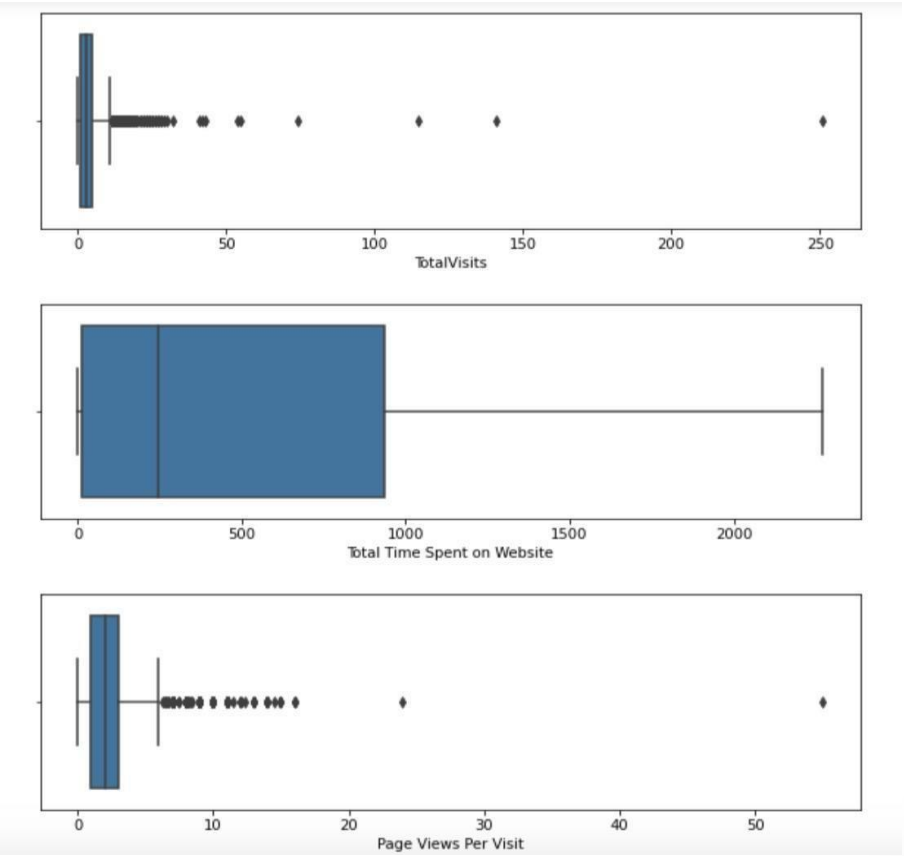
DATA SOURCING, CLEANING AND PREPARATION

- Python libraries were used for the same.
- Data was checked for duplicate entries and none were found.
- Missing values in the data were dealt with using various methods like

    - Replacing the missing values with mode for categorical data and median or mean for numerical data.
    - Columns with very high percentage of data which did not have much relevance were dropped

DATA SOURCING, CLEANING AND PREPARATION

- Outlier treatment of data was done on the basis of percentiles of the data range.
- Boxplot was made to see this graphically.
- It can be seen that outliers exist in the columns TotalVisits and Page Views Per Visit columns.



DATA SOURCING, CLEANING AND PREPARATION

- Some outlier data is deleted which may skew our results by making our model less accurate
- Some columns which will not contribute to our analysis have been dropped
- Example - What matters most to you in choosing a course, columns that have more than 30% null values have been dropped…..
- After cleaning data 98% data has been retained
- There are 9090 entries and 27 columns

EDA

UNIVARIATE ANALYSIS

BIVARIATE ANALYSIS

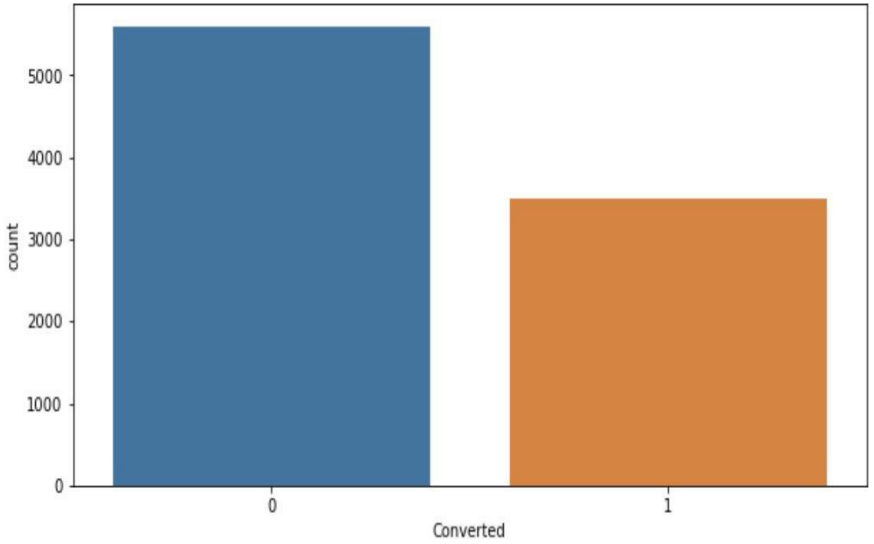
MULTIVARIATE ANALYSIS

EDA - UNIVARIATE ANALYSIS

CONVERTED VARIABLE –

This is also our target variable. Indicates whether a lead has been successfully converted or not

The count of leads converted is almost half of the leads that are not converted.



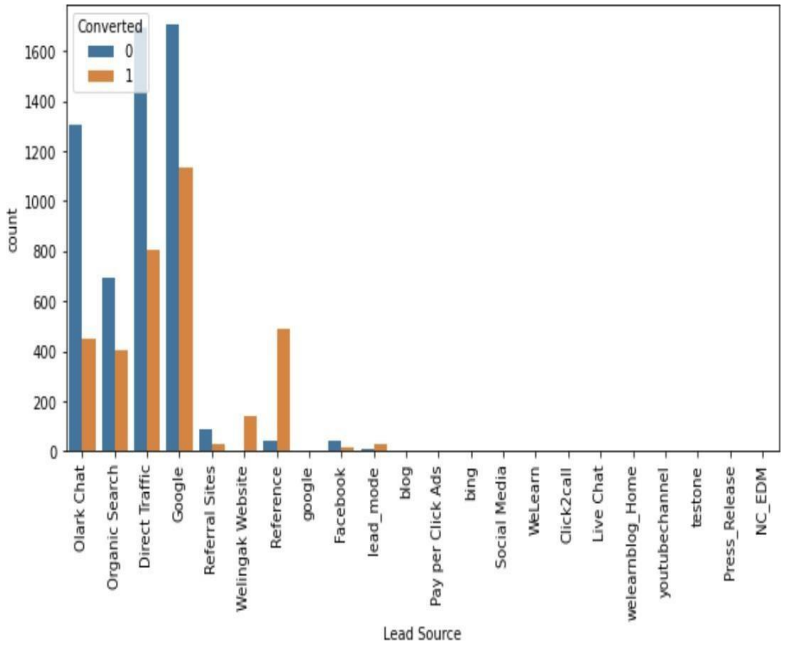
Note - We can clearly conclude that most of leads are not successfuly converted.

EDA - BIVARIATE ANALYSIS

LEAD SOURCE &CONVERTED VARIABLE -

We plotted the Converted variable against the Lead Source to determine which Sources contribute most to the Converted leads

Traffic from Google searches have the most conversions

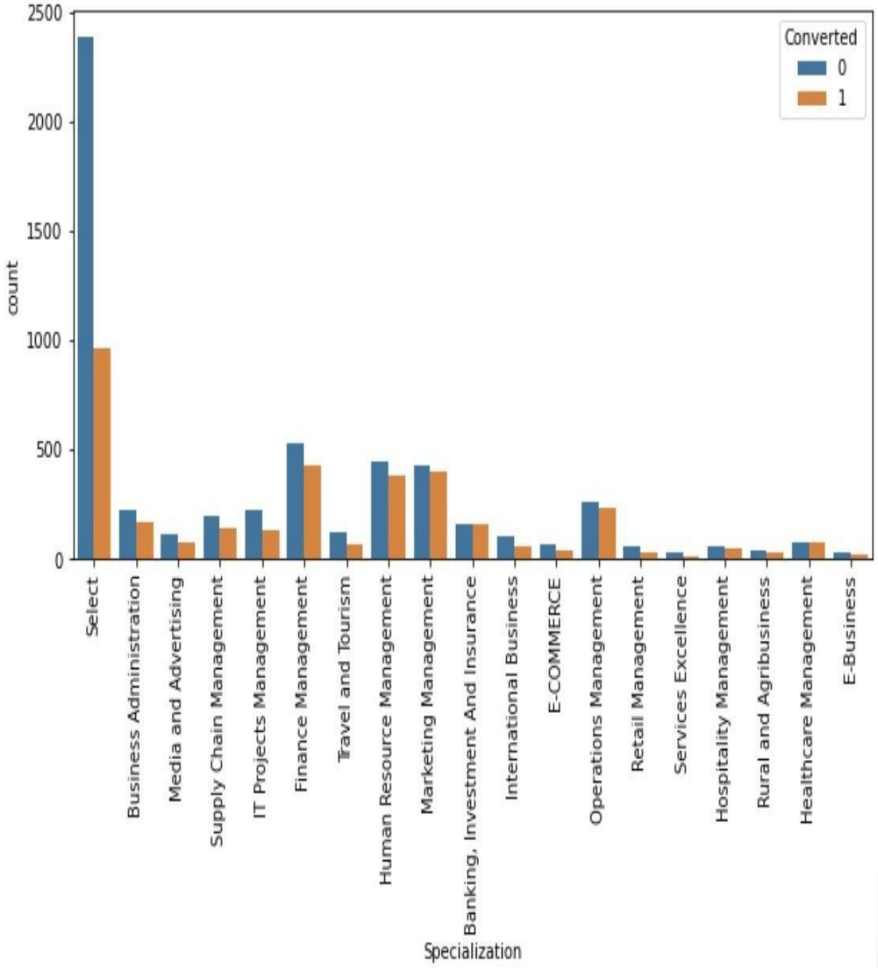


EDA - BIVARIATE ANALYSIS

SPECIALIZATION & CONVERTED VARIABLE –

We plotted the Converted variable against the Specialization to determine which subject area contribute most to the Converted leads.

Most leads have not mentioned their area. Out of others Finance has a good number of conversions.

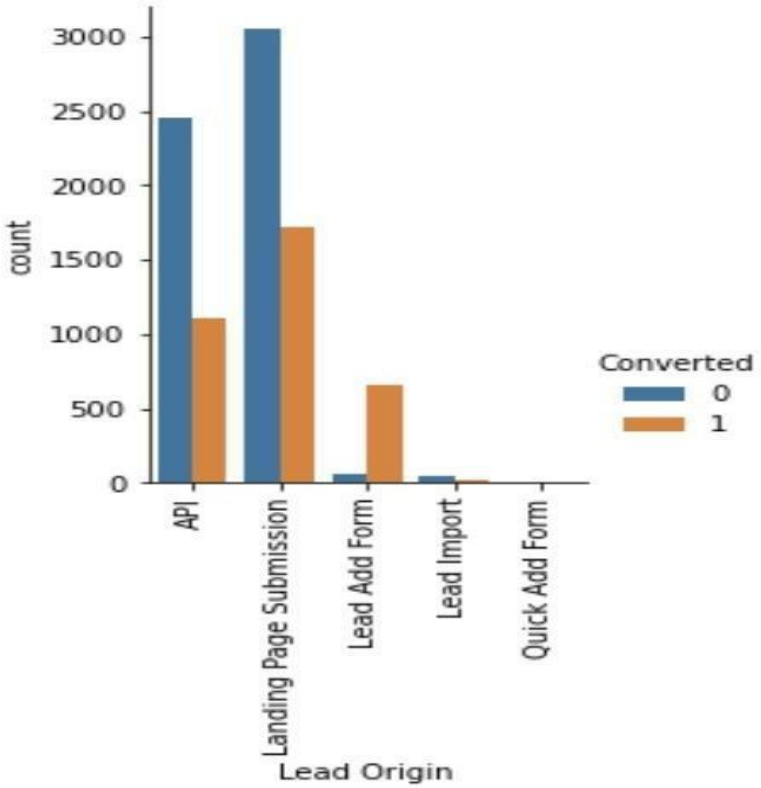


EDA - BIVARIATE ANALYSIS

LEAD ORIGIN & CONVERTED VARIABLE :

We plotted the Converted variable against the Lead Origin to determine which subject area contribute most to the Converted leads

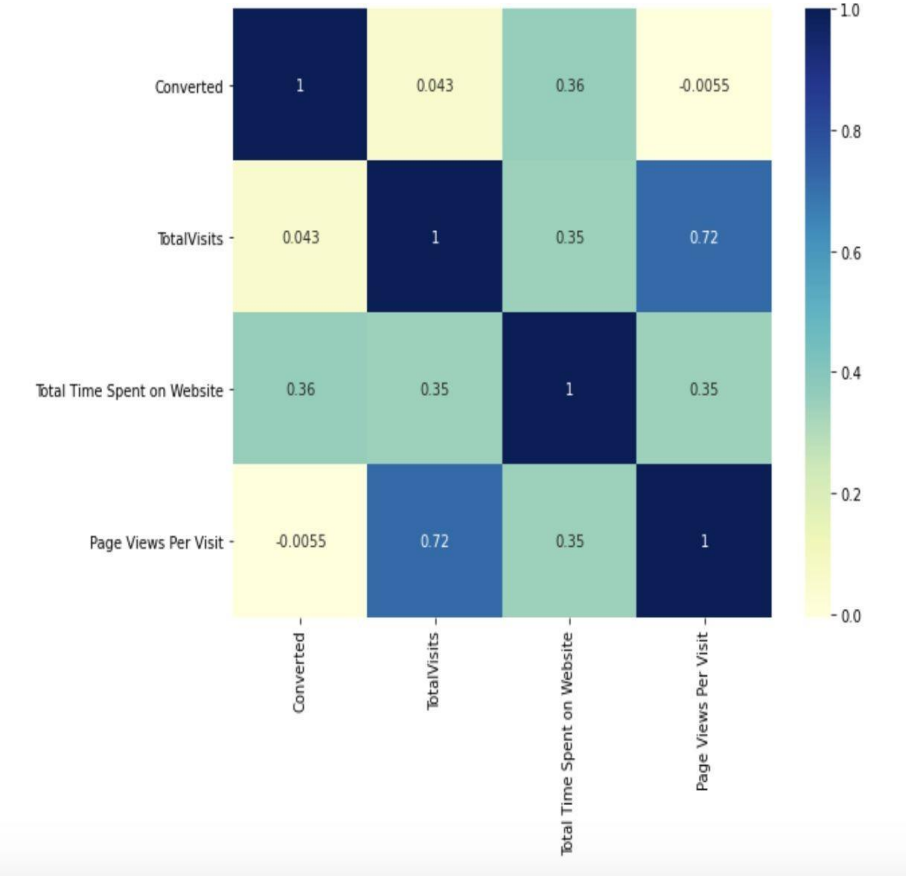
Leads from the Landing Page Submission have a good conversion rate



EDA - MULTIVARIATE ANALYSIS

Pairplots and heatmaps were created to understand the correlation between the numerical variables.

- **Total Visits and Page Views Per Visit has a strong correlation. In such a case, one may be dropped before analysis starts.**



DATA CLEANING BASED ON EDA RESULTS

- Based on the EDA analysis it is seen that many columns are not adding any information to the model, hence we can drop them before further analysis.
- The following columns were dropped before model building -
  - 'Do Not Email', 'Do Not Call','Search', 'Magazine', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations','Receive More Updates About Our Courses', 'Update me on Supply Chain Content', 'Get updates on DM Content','Tags', 'I agree to pay the amount through cheque'

DATA CLEANING BASED ON EDA RESULTS

- For categorical variables with multiple levels, dummy features - variables are created using **one-hot encoding**
- Variable used in this part are -

  'Lead Origin', 'Lead Source', 'Last Activity', 'Country', 'Specialization', 'How did you hear about X Education', 'What is your current occupation', 'A free copy of Mastering The Interview', 'Last Notable Activity'
- After the dummy variables are created, these original variables are dropped

DATA CLEANING BASED ON EDA RESULTS

- The top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion are -
  - Lead Origin Lead Add Form
  - Last Notable Activity SMS Sent
  - What is your current occupation Working Profession

MODEL BUILDING

- For the system at hand, we have built a **LOGISTIC REGRESSION MODEL.**
- The data at this stage is divided into -
  - Train Data (80%)
  - Test Data (20%)
- The target variable is separated into a different dataframe.
- We have used RFE for feature selection.
- We have received the top 18 features of the model.

MODEL BUILDING

- Feature Scaling has been performed on the following variables -
  - 'TotalVisits'
  - 'Total Time Spent on Website'
  - 'Page Views Per Visit'
- Feature Selection is done using the RFE method and the number of variables is reduced to 18. Then manually the model summary is analyzed and variables having a p-value of more than 0.05, are dropped one by one. After dropping each variable, the model summary is again analyzed.
- Optimal Cut-Off Point is calculated and comes out to be 0.4

MODEL BUILDING - LEAD SCORE

- Main aim was to calculate a Lead Score between 0 - 100 for each Lead in the dataset.
- We are using the Converted Probability Score to calculate the Lead Score.
- Sample scores can be seen here -

	Converted	Converted_Prob	final_predicted	lead_score
3343	1	0.278202	0	28
2600	1	0.567468	1	57
3115	0	0.074880	0	7
1348	0	0.031186	0	3
4525	1	0.935047	1	94
461	0	0.052770	0	5
1079	1	0.939143	1	94
7339	0	0.031186	0	3
6291	0	0.120845	0	12
1672	0	0.081303	0	8

MODEL EVALUATION, COMPARISON & CONCLUSION

- Accuracy, Sensitivity and Specificity values of test set are around 75%, 78% and 82% which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%
- Hence, overall this model seems to be good.

BUSINESS RECOMMENDATIONS - 1

The top three variables in model which contribute most towards the probability of a lead getting converted are

- **Total Time Spent on Website**
- **Lead Origin Lead Add Form**
- **Last Notable Activity SMS Sent**

BUSINESS RECOMMENDATIONS - 2

Aggressive workflow for converting leads :

- High sensitivity implies that our model will correctly identify almost all leads who are likely to convert
- It will do that by overestimating the Conversion likelihood
- To follow an aggressive workflow choose a lower threshold value for Conversion Probability.

  This will ensure the Sensitivity rating is very high which in turn will make sure almost all leads who are likely to Convert are identified correctly and the agents can make phone calls to as much of such people as possible

BUSINESS RECOMMENDATIONS - 3

Whom to approach :

- People who spend more time on the website and this can be done by making the website more informative and thus bringing them back to the site
- People who repeatedly visit the website
- Customers whose last activity is through SMS or through Olark chat conversation
- Customers who are working professionals

Thank You