



# Lead Scoring

*CASE STUDY RESULTS BY: MEGHNA RATHI*

# Business Understanding

- ▶ In education industry's online courses provision domain, the companies selling various online courses to industry professionals want to work to improve their lead conversion rate, wherein target customers need to be identified who will convert i.e., sign up for any of the online courses
- ▶ For such companies, say they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the online courses selling 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
  - ❑ There are a lot of leads generated in the initial stage but only a few of them come out as paying customers from the bottom. One needs to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc. ) in order to get a higher lead conversion

# Problem Statement

- ▶ An education company, named 'XEducation', sells online courses to industry professionals who are interested in their courses - on any given day, many people land on their website and browse for courses
- ▶ The company wants us to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers
- ▶ The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given us a ballpark of the target lead conversion rate to be around 80%

## Objective

- ▶ The company wants us to identify the most potential and promising leads, also known as '**Hot Leads**', such that the lead conversion rate is good for the company and the leads that are most likely to convert turn into paying customers

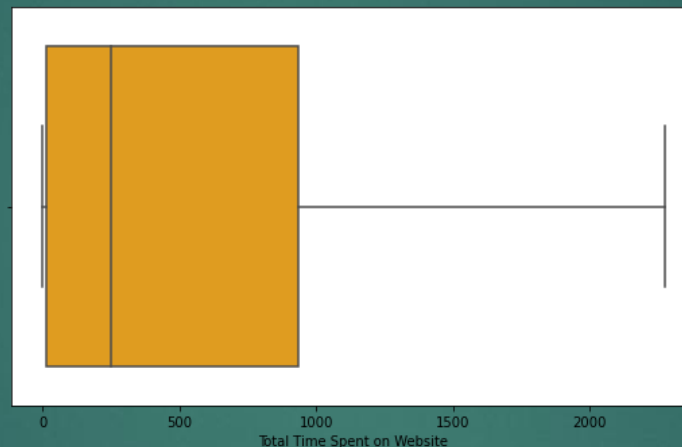
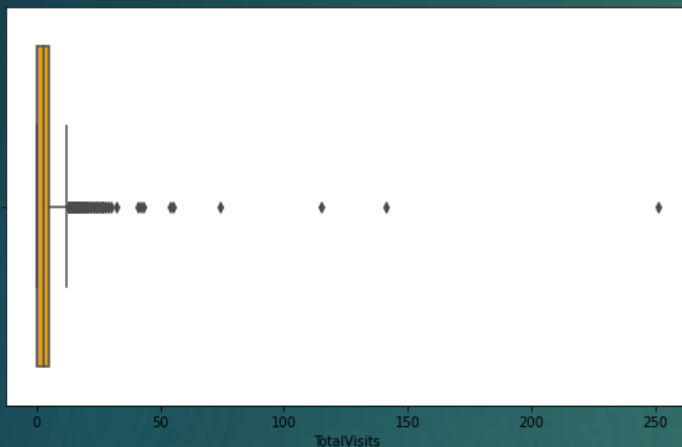
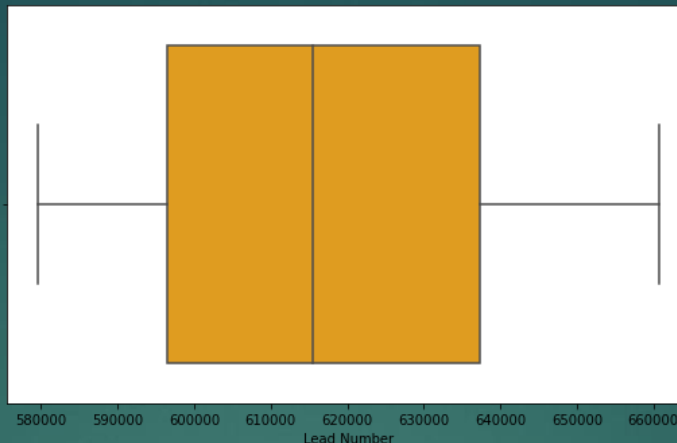
# Assumptions

- ▶ Data values such as 'Select' and blanks constitute as NULLs
- ▶ Leads data columns with missing value proportions > 30% are to be dropped (few exceptions) as insufficient data to perform analysis or missing value imputation, without introducing bias/irregularities in data
- ▶ Blanks in categorical data replaced using 'Others/Mode'
- ▶ Dropped skewed columns where the data is limited to a particular category as they won't be effective for analysis
- ▶ Attributes having >90% values as "No" are dropped as not important for analysis
- ▶ Dropped 'Last Activity' and 'Last Notable Activity' columns as we want only customer based inputs for analysis
- ▶ Dropping 'TotalVisits' more than 10 (95 percentile) as they are very high compared to mean, so treated as suspicious/erroneous data

# Overall Approach: Step-by-Step

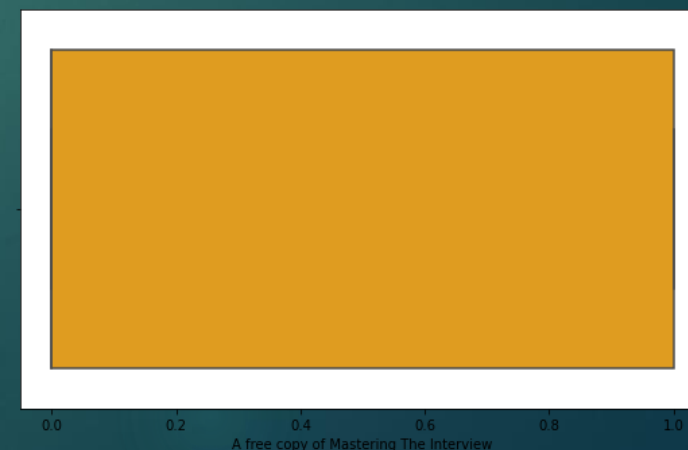
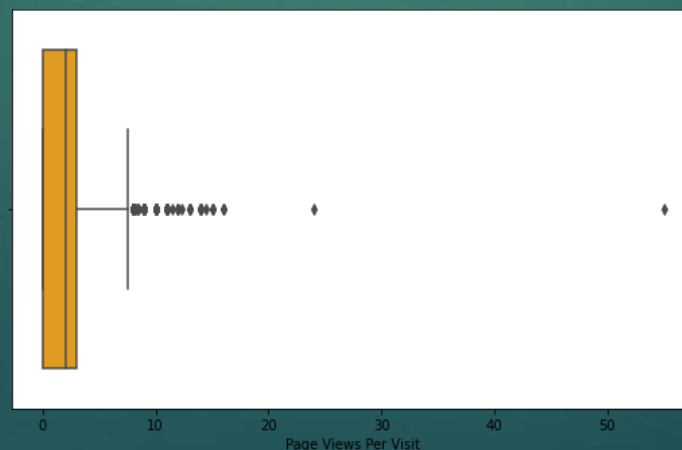
- ▶ Sourcing the datasets and understanding them via descriptive information of all columns
- ▶ Performing data quality and missing value checks, along with missing value imputation as part of establishing data sanity
- ▶ Deep-dive into data using varied univariate, bivariate, multivariate visualizations and extensive data cleaning, variable transformations and treating the outliers
- ▶ Checking for data imbalance and correlations
- ▶ Variables conversion into appropriate formats – readying the data for logistic model building via dummy variables' creation, train-test data splitting and feature scaling
- ▶ Model building (iteratively), finding optimal cut-off using ROC curve, testing model on test data and model evaluation
- ▶ Generated the lead score variable with respective lead numbers where a high lead score indicated a hot lead

# Univariate Analysis



- Based on this, dropping 'TotalVisits' more than 10 (95 percentile value) as they are very high compared to the mean value of 3.39

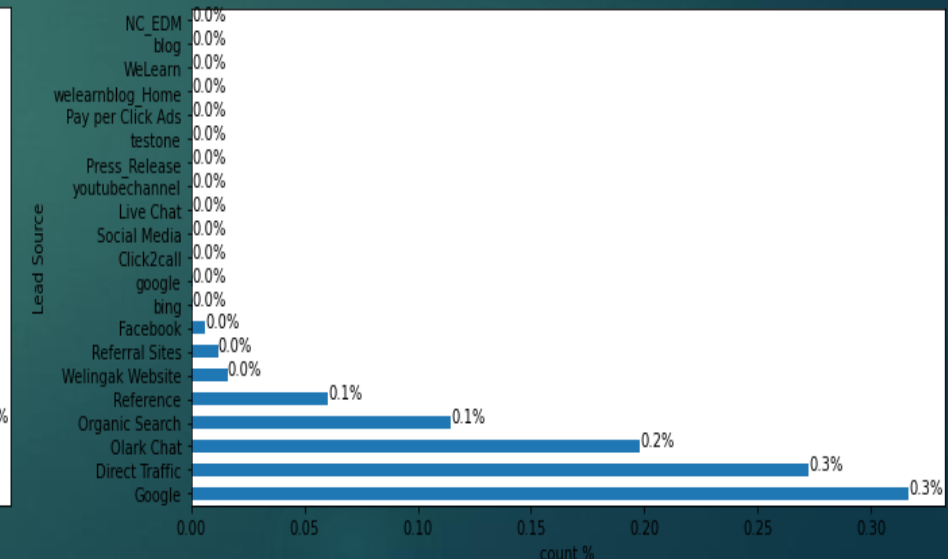
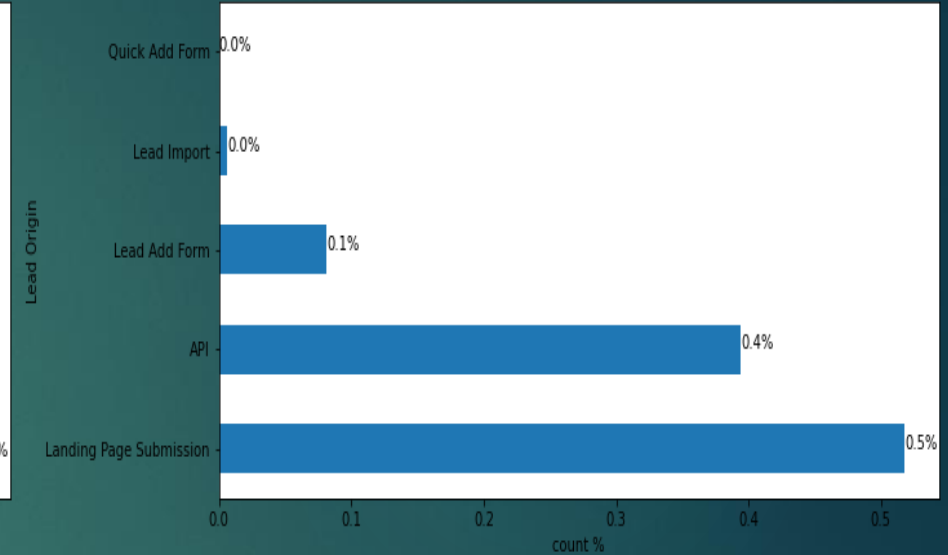
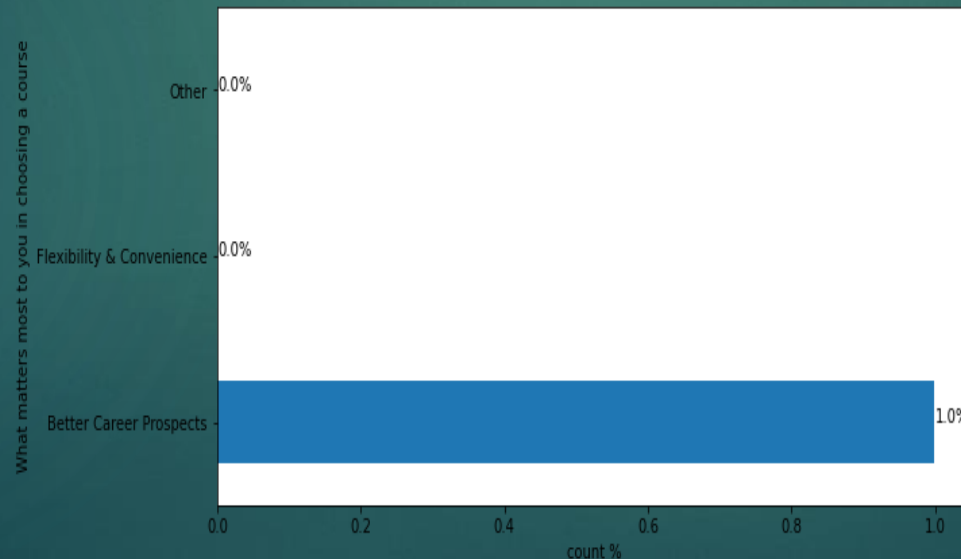
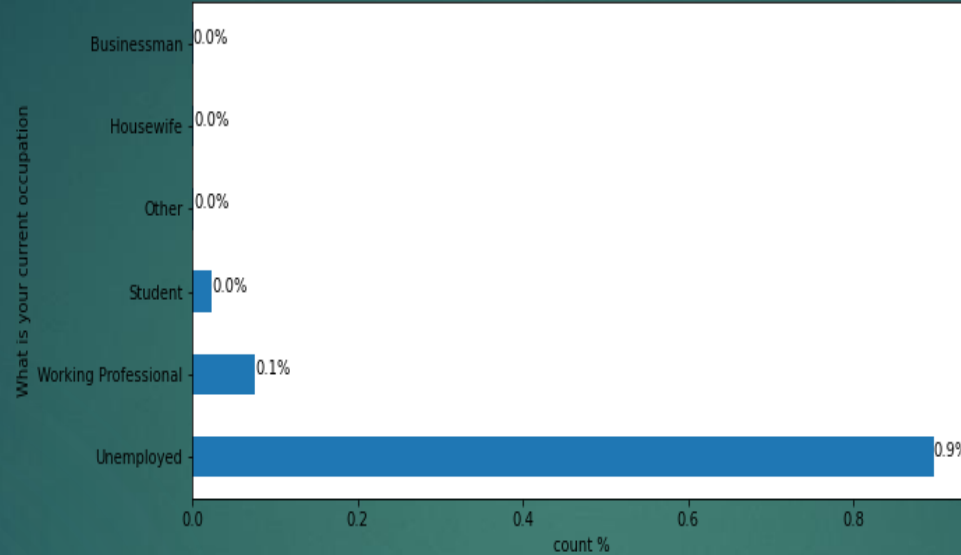
- Outliers in 'Page Views Per Visit' also got handled as per outlier treatment for 'TotalVisits' variable





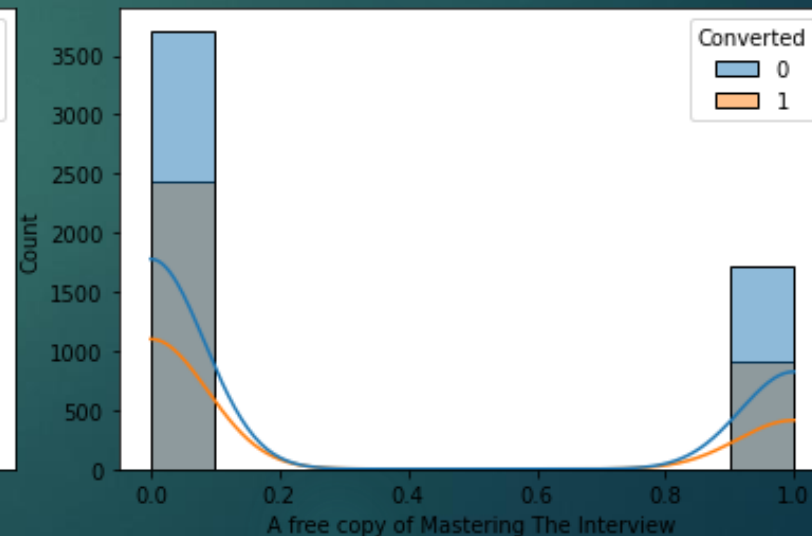
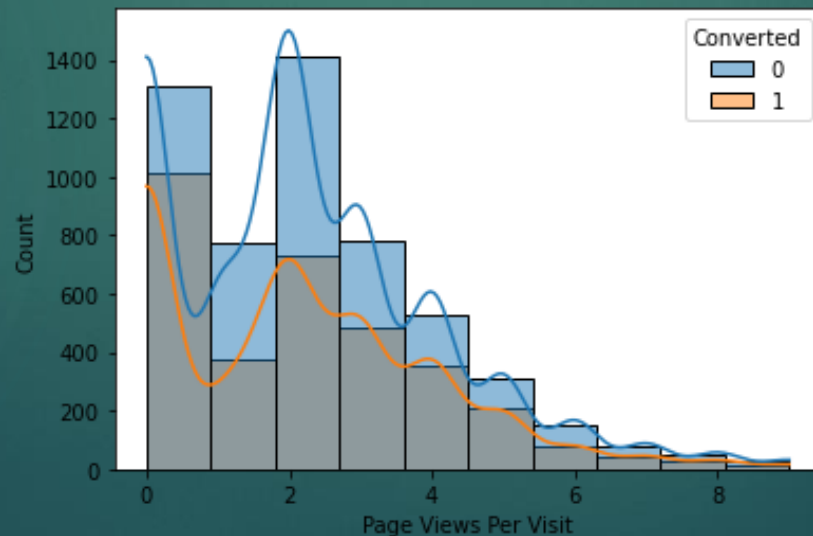
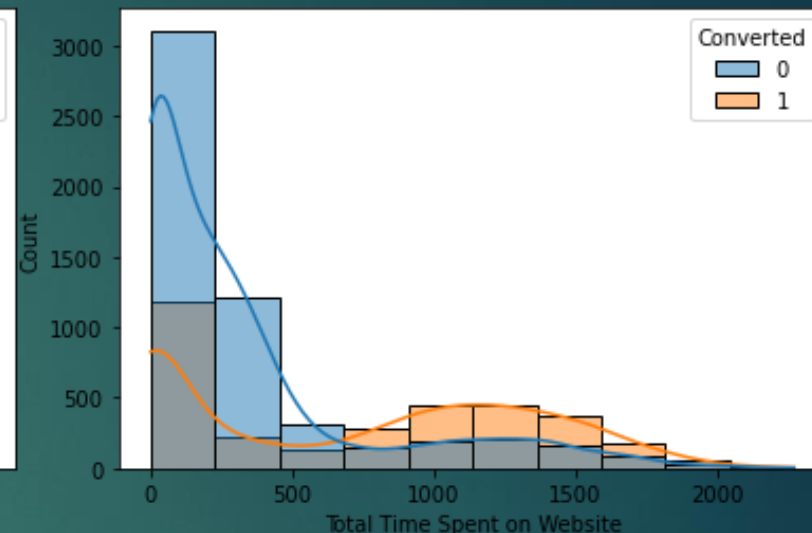
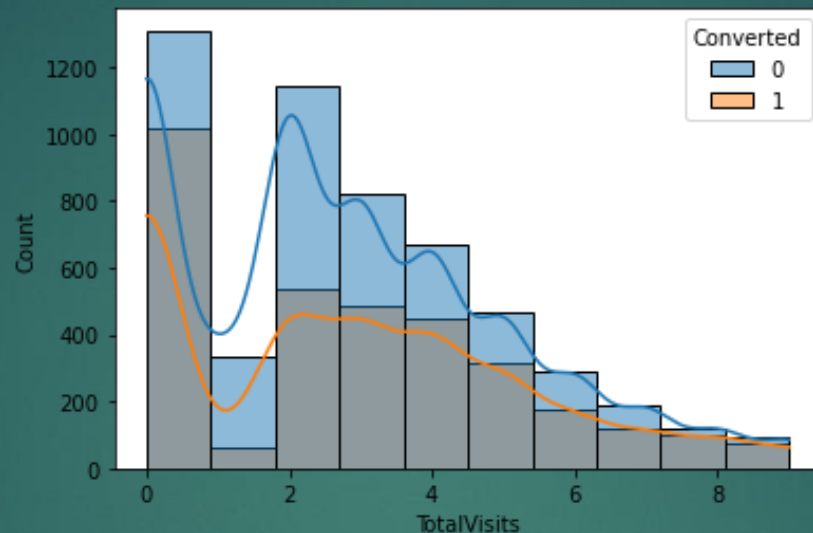
# Univariate Analysis

- Based on this, performed grouping the similar values/categories in single bucket within the variables like 'Lead Source' and 'Lead Origin' and renamed columns to shorter versions



# Bivariate Analysis

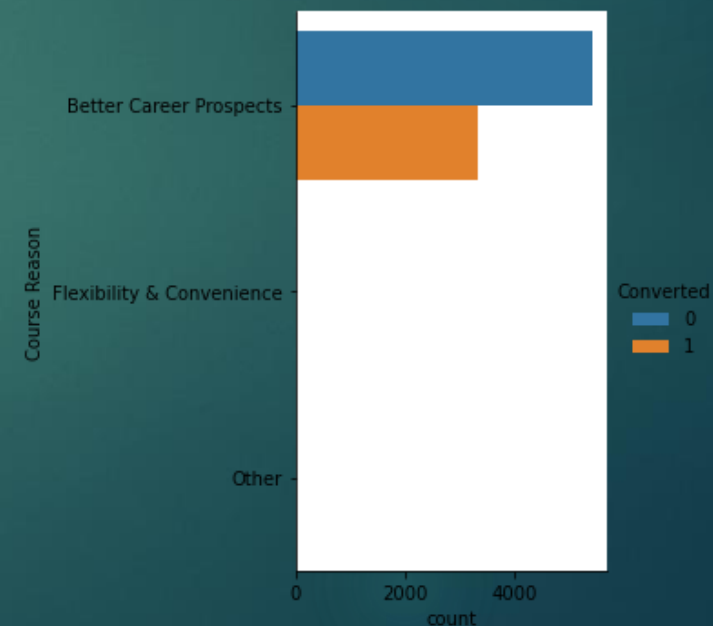
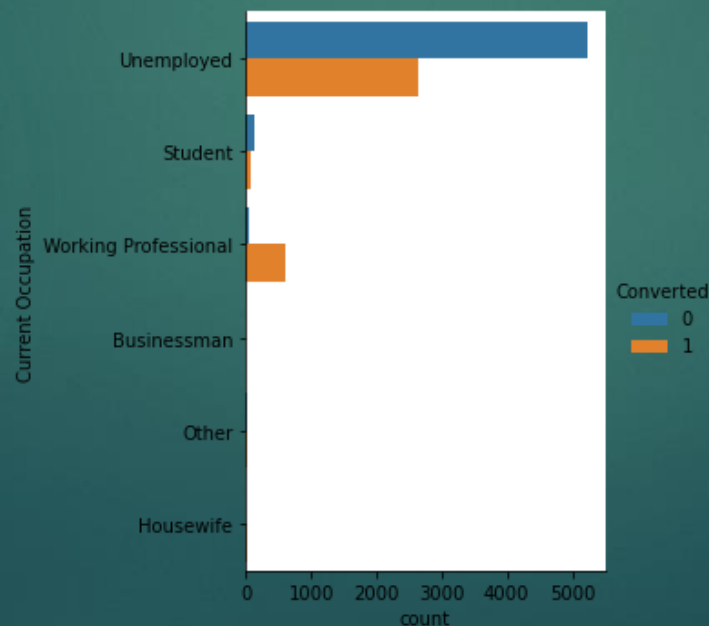
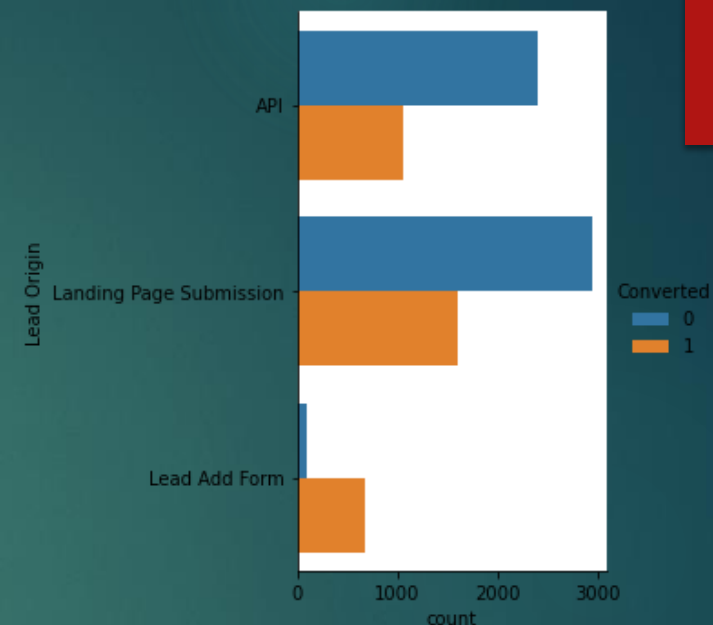
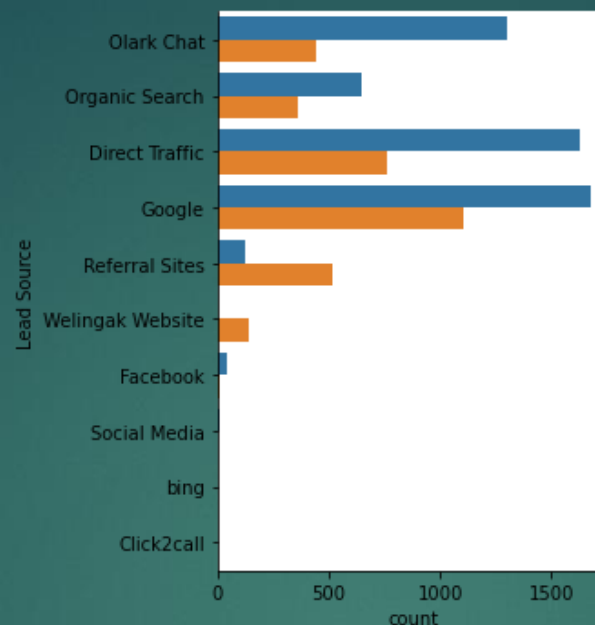
- ▶ If 'Total Time Spent on Website' is in good range, we have good lead conversion rate
- ▶ As 'Total Visits' and 'Page views Per Visit' increase, the conversion seems to lessen indicating weak or negative correlation with target variable
- ▶ Free copy of interview class inclusion needs to be looked at to identify relationship with target variable





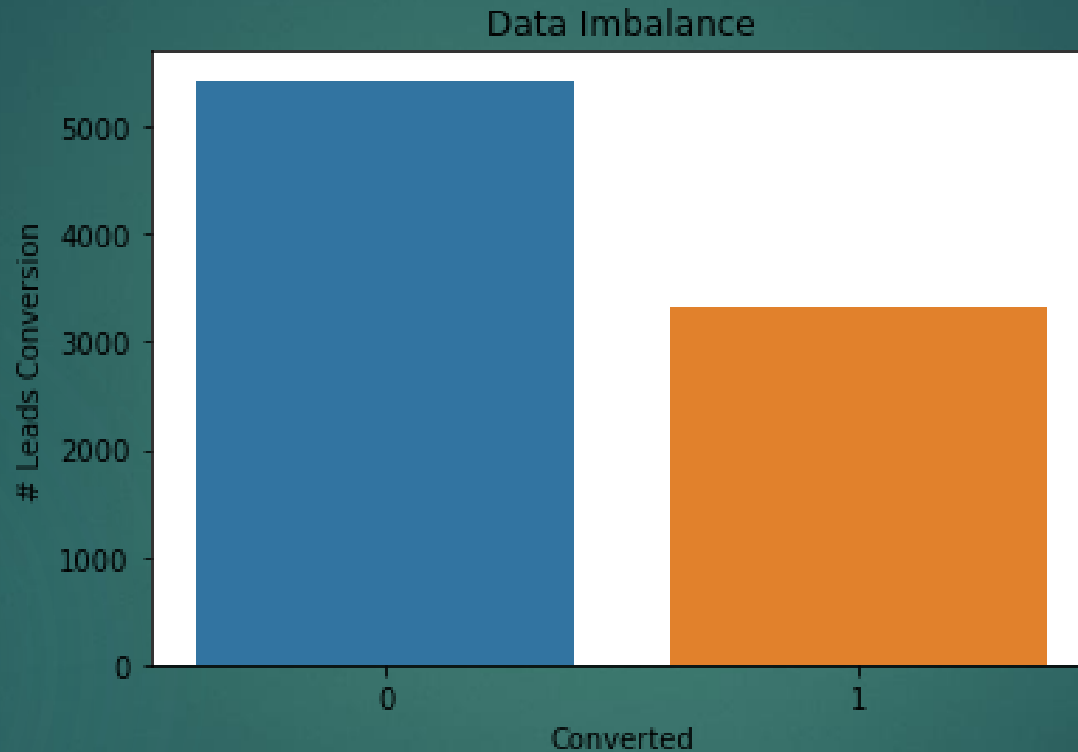
# Bivariate Analysis

- ▶ 'Google', followed closely by 'Direct Traffic' are having highest conversion among all Lead Sources
- ▶ People want 'Better Career Prospects' as the outcome for the courses to be considered a good fit for them
- ▶ 'Unemployed' people have very bad conversion rate whereas 'Working Professionals' have good conversion rate - 'Current Occupation' is going to be a good indicator for identifying hot leads
- ▶ Lead Origin looks an okay indicator for hot leads identification - correlation needs to be looked at to identify relationship with target variable



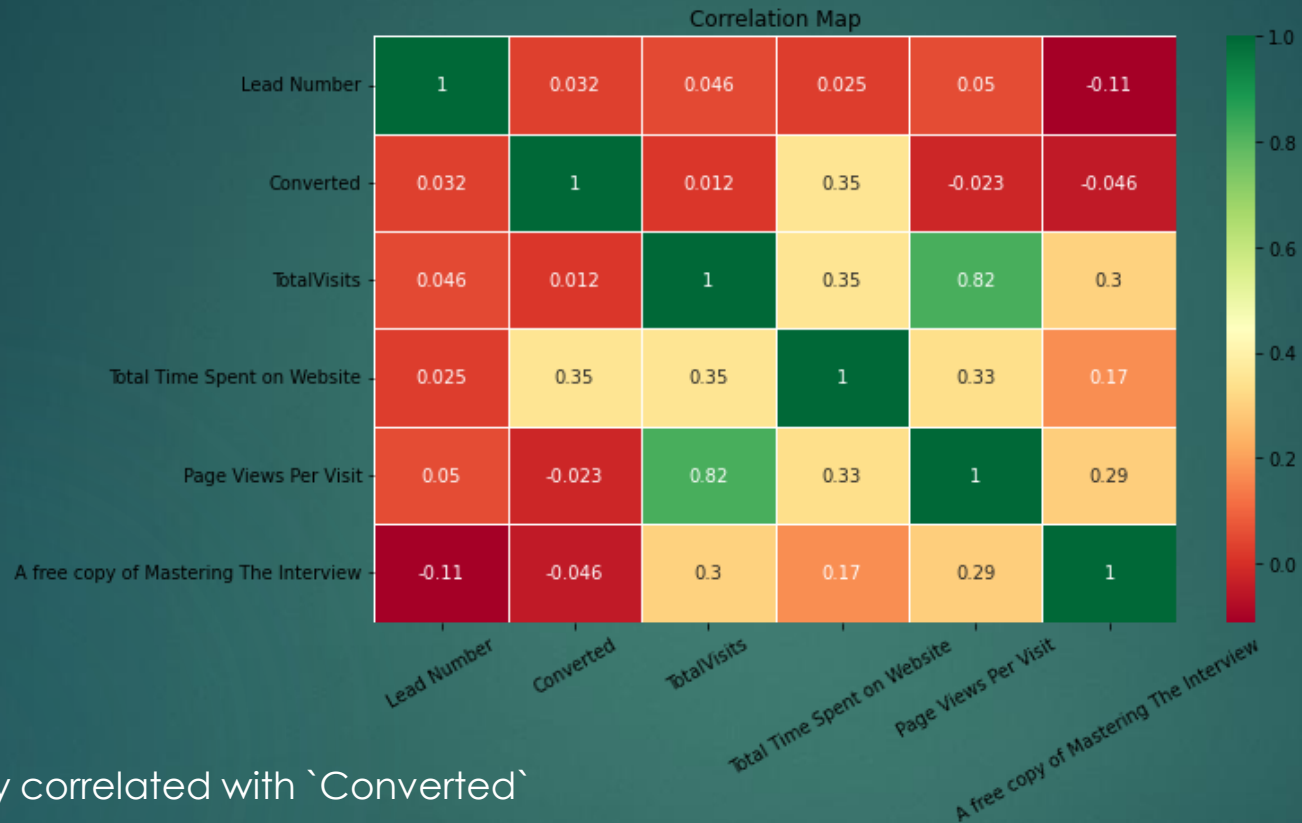
# Data Imbalance

Non-converts: 61.9% Converts: 38.1% Imbalance Ratio: 0.61



- ▶ If there is a greater imbalance ratio, the output is biased to the class which has a higher number of examples
- ▶ Since, our leads data has Imbalance Ratio of 0.61, which is very less (negligible), our output will not be biased to class of non converted (as non-converters class have higher count of examples). This is good for our analysis.

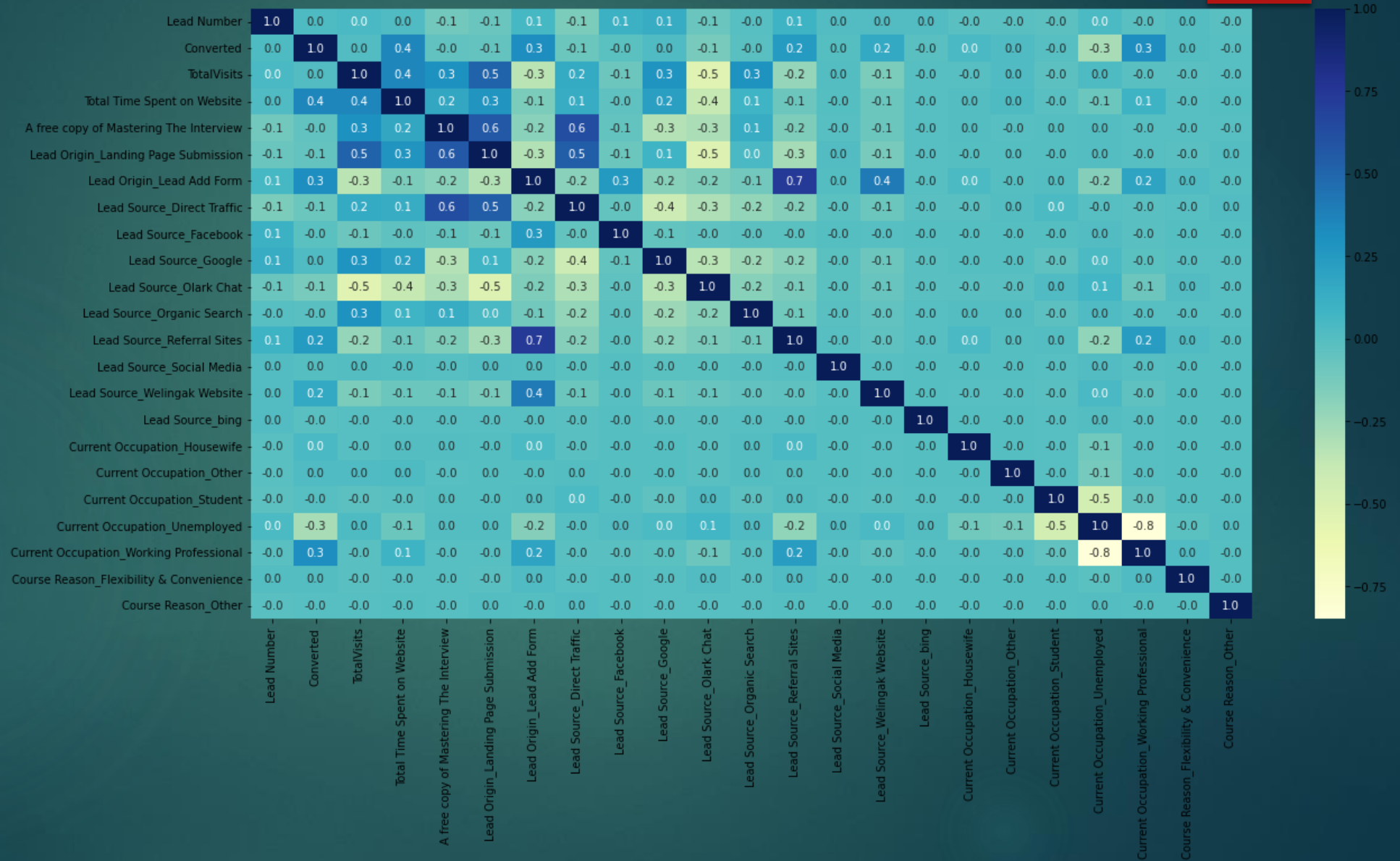
# Correlation Heatmap



- ▶ `TotalVisits` is weakly correlated with `Converted`
- ▶ `Page Views Per Visit` is negatively correlated with `Converted` target variable
- ▶ This is very highly correlated with 0.82 correlation with `TotalVisits` and we'll be dropping this variable as information of page visits is captured in the total variable itself
- ▶ `Total Time Spent on Website` is having moderate correlation with `Converted` target variable
- ▶ 'Free copy of Mastering Interview' is highest correlated with `Converted` for leads amongst all numeric predictors - shows its inclusion might be important in converting the leads



# Correlation Matrix

- ▶ We see that few variables are highly correlated with each other, which might result in multicollinearity
- ▶ We shall address this after p-value elimination and final VIF check while building the model in iterative process



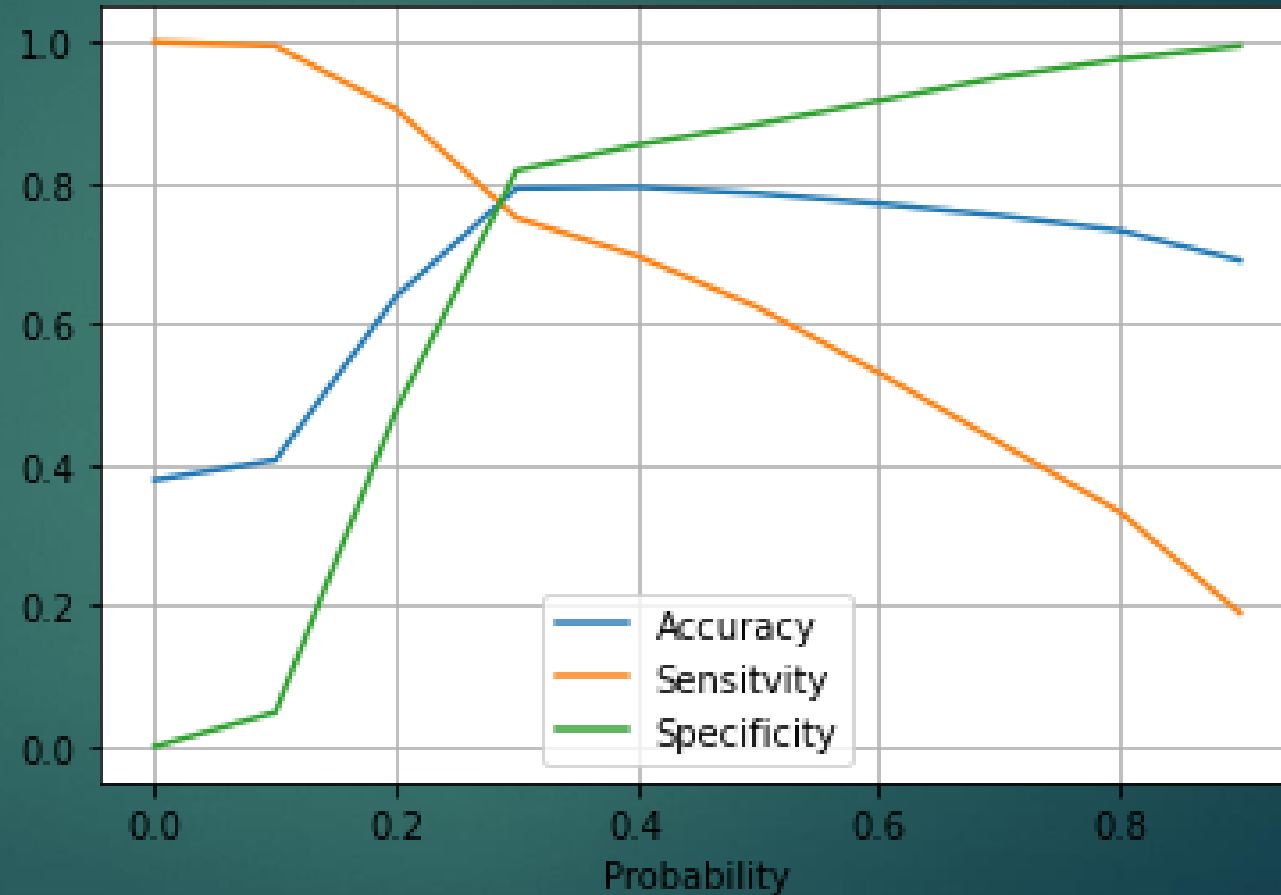
# Confusion Matrix

- ▶ As per our business objective of identifying all hot leads accurately (and not missing out on any promising lead), the recall (or sensitivity) score is more valuable because it is okay if our precision is little low which means lesser hot leads, but we don't want to miss out on any hot leads who are willing to get converted, thereby increasing the conversion rate (our main goal). Hence, our focus will be more on 'Recall' than Precision
- ▶ Metric values are as follows (on test dataset):
  - ❑ Precision: 71.6%
  - ❑ Recall: 76%
  - ❑ Accuracy: 79.3%

<b>Actual Predicted</b>  	<b>Negative</b>	<b>Positive</b>
<b>Negative</b>	3,121	696
<b>Positive</b>	580	1,744

# Optimal Threshold: Cut-off Value

- From the given curve, we can see that nearly 0.3 is the optimum point for considering probability cutoff value for getting predictions on our leads data





# Model summary

- ▶ We can see that we after our model build, the company's sales team will be able to identify **79.3% of hot leads** accurately with a **recall score (sensitivity) of 76.5%** - this will enable the company to attain a higher conversion rate based on a good identification of promising leads
  - ❑ We have arrived at a good model for the hot leads identification with the significant variables
- ▶ Train and Test model - Metrics comparison:
  - ❑ Training Accuracy: 79.2%
  - ❑ Testing Accuracy: 79.3%
- ▶ The top 3 features for identifying most promising hot leads, thereby ensuring a good conversion rate, are:
  - ❑ Lead Origin\_Lead Add Form with the coefficient of 4.18
  - ❑ Current Occupation\_Working Professional with the coefficient of 2.719
  - ❑ Lead Source\_Welingak Website with the coefficient of 2.024

# Recommendations

- ▶ 'Lead Add Form' as origin for lead customers identifier is the best. So, XEducation can do more promotions on this lead origin identifier and focus on customer targeting using the same
- ▶ XEducation's sales team should focus more on leads whose current occupation is 'Working Professional' as they represent good target base for lead conversion
- ▶ 'Welingak Website' as the Lead source is very effective for getting promising leads, so XEducation can provide some promotional discounts to increase traffic and lead conversion using this source.
  - ❑ Also, Olark Chat is a good lead source as well, so sales team can put resources on this as well
- ▶ 'Lead Source\_Facebook' is having a negative coefficient of -0.3441, so Sales team should look into this lead source mode and not put more investment on it