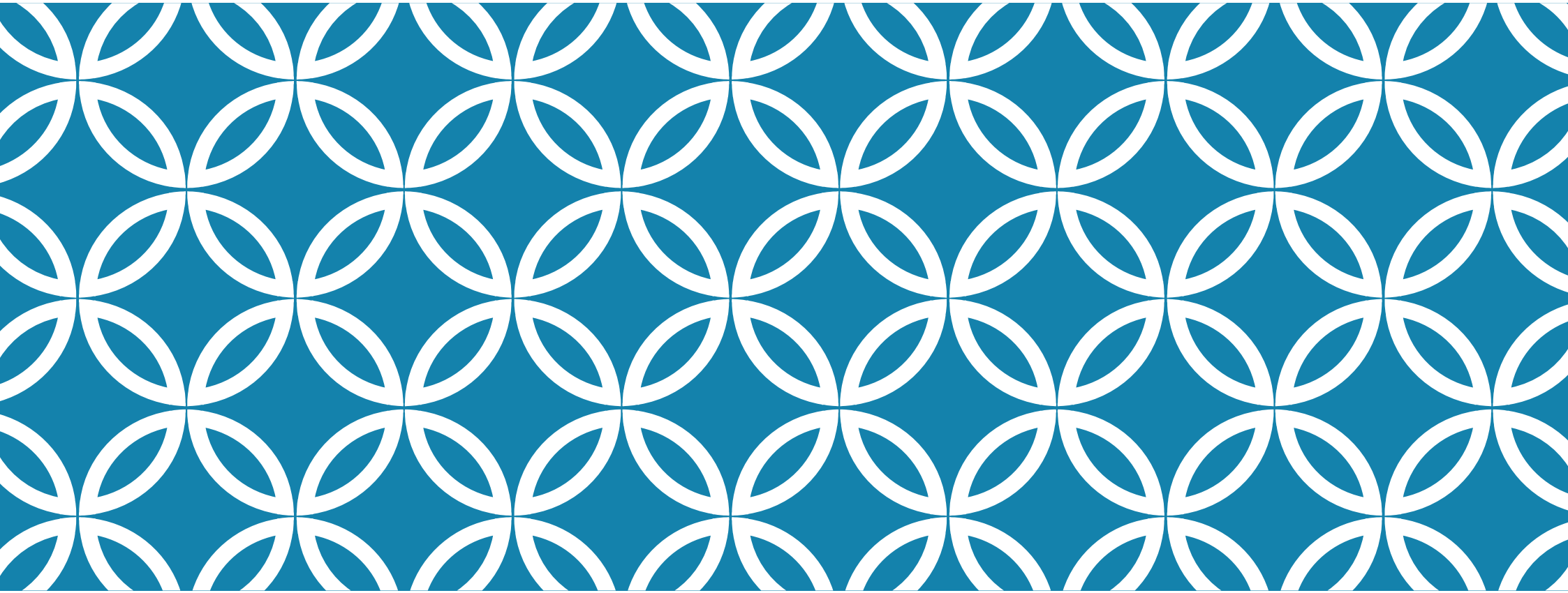


# LEAD SCORING CASE STUDY

By Abhishek Bose

# PROBLEM STATEMENT

An education company named X Education sells online courses to industry professionals.

The company markets its courses on several websites & search engines like Google. Once these people I& on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals.

Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

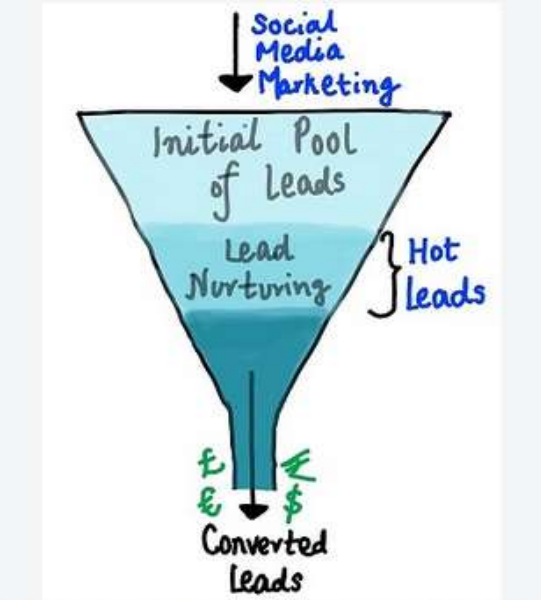
Now, although X Education gets a lot of leads, its lead conversion rate is very poor. Around 30% of leads in a day are converted.

# OBJECTIVE

X Education needs help in selecting the most potential leads also known as Hot Leads which are the leads that are most likely to convert into paying customers.

The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance & the customers with lower lead score have a lower conversion chance.

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.



Lead Conversion Process - Demonstrated as a funnel

# ANALYSIS APPROACH

- Data cleaning & data manipulation
  - Check & replace duplicate data (Select).
  - Check & handle NaN values & null values.
  - Drop columns, if it contains large amount of missing values & not useful for the analysis.
- Imputation of the values.
- Check & handle outliers in data.
- EDA
  - Bivariate data analysis: correlation coefficients & pattern between the variables etc.
  - Univariate data analysis: value count, distribution of variable etc.
- Feature Scaling & Dummy Variables & encoding of the data.
- Using Logistic Regression for the model making & prediction.
- Validation of the model using classification report.

# DATA MANIPULATION

We need to check the nulls in the data, additionally as per data dictionary provided "Select" indicates customer has not selected the specialisation while filing the form which can be considered as null value.

We will outright drop the columns with >50% of null values & the columns which we won't use in analysis as per the data dictionary such as:

- Lead Profile : >50% null or 'Select'
- Lead Quality : >50% null or 'Select'
- Prospect ID : unique ID cant be used for analysis, only checked for duplicates
- Lead Number : unique ID cant be used for analysis
- Asymmetrique Profile Score : >49% null
- Asymmetrique Activity Score : >49% null
- Asymmetrique Activity Index : >49% null
- Asymmetrique Profile Index : >49% null
- Tags : No use for analysis
- Last Notable Activity : No use for analysis
- Last Activity : No use for analysis
- How did you hear about X Education : >50% null or 'Select'

# DATA MANIPULATION

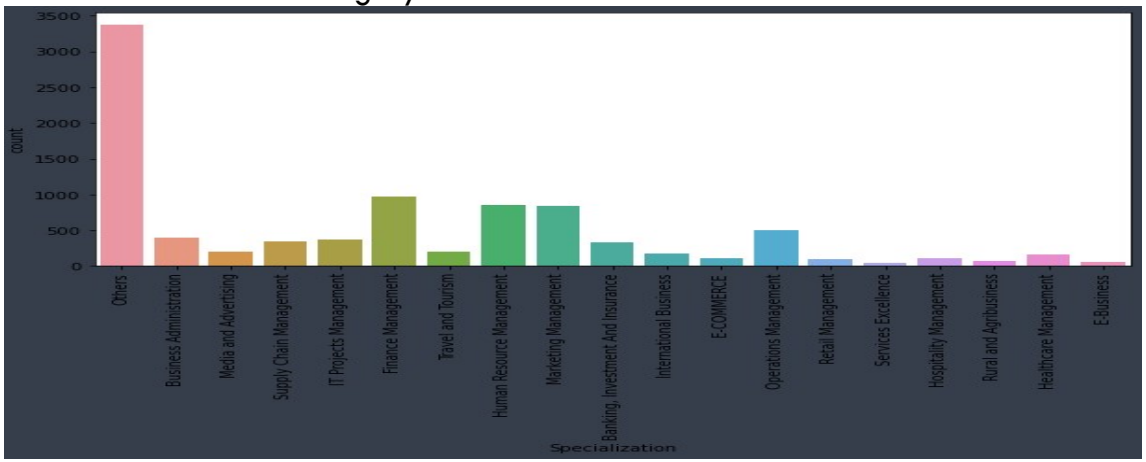
Following columns only have 1 option selected in each, it can be said they have no affect on the lead scoring:

- 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

Null values in City columns could be imputed based on country values, however as country values were not related to cities Ex. Country: Australia City: Mumbai we will drop the City column and use the country column

For imputation of null values of Specialisation column we will check value counts

As the blank cases might be actual students or recent graduates looking for jobs which do not fall in any specialisation
Hence we cannot impute the values with mode of the column We should create a category “Others” for all such cases




```
leads.Specialization.value_counts(normalize = True)*100
```

| | |
|---|---|
| Finance Management | 16.655290 |
| Human Resource Management | 14.470990 |
| Marketing Management | 14.300341 |
| Operations Management | 8.583618 |
| Business Administration | 6.877133 |
| IT Projects Management | 6.245734 |
| Supply Chain Management | 5.955631 |
| Banking, Investment And Insurance | 5.767918 |
| Travel and Tourism | 3.464164 |
| Media and Advertising | 3.464164 |
| International Business | 3.037543 |
| Healthcare Management | 2.713311 |
| Hospitality Management | 1.945392 |
| E-COMMERCE | 1.911263 |
| Retail Management | 1.706485 |
| Rural and Agribusiness | 1.245734 |
| E-Business | 0.972696 |
| Services Excellence | 0.682594 |

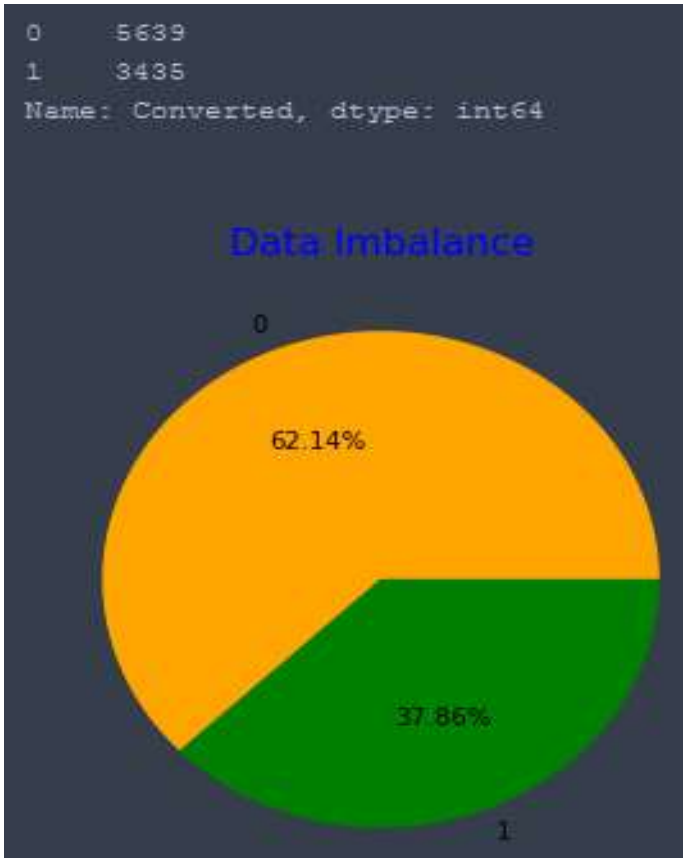
Name: Specialization, dtype: float64

# DATA MANIPULATION

- For “What matters most to you in choosing a course” column : Being a education company it is obvious "Better Career Prospects" would be the answer by most of the customers which is indicated by highly skewed data, it is futile to impute the missing values rather we should drop the column altogether
- For “What is your current occupation” column the clients have intentionally left it blank instead of selecting Others, we will impute the missing value to “Unknown”
- For “Country” column the data is highly skewed to India it should be dropped
- In columns “Page Views Per Visit” ,” TotalVisits”, “Lead Source” the missing values are very miniscule it is more preferable to drop these columns
- Finally the columns are renamed to more appropriate names
  - What is your current occupation = Occupation
  - Through Recommendations = Recommendations
  - A free copy of Mastering The Interview = Free Copy

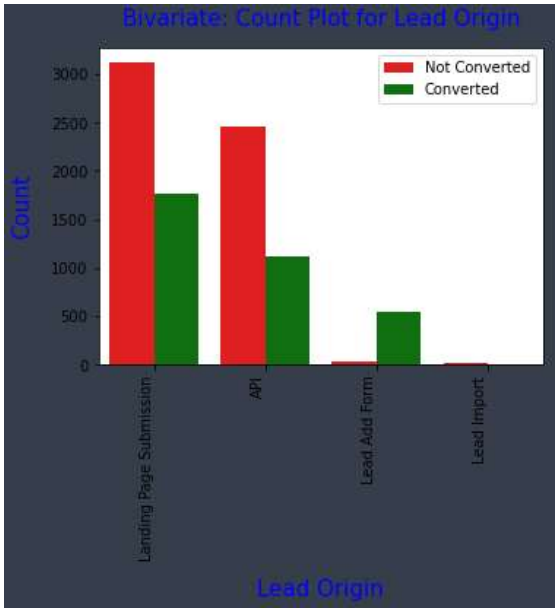
# DATA IMBALANCE

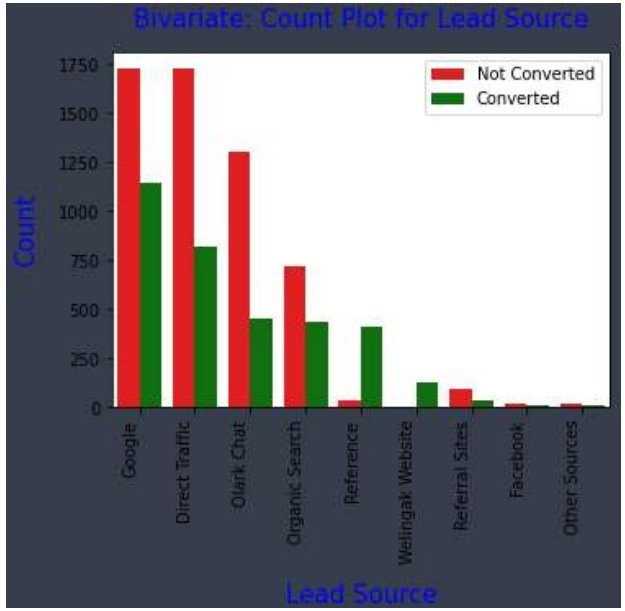
In the lead conversion ration, 37.86% has converted to leads where as 62.14% did not convert to a lead.
So it seems like a almost balanced dataset.



# BIVARIATE ANALYSIS –CATEGORICAL COLUMNS

To improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.
To increase lead count, initiatives should be taken so already existing members increase their referrals.



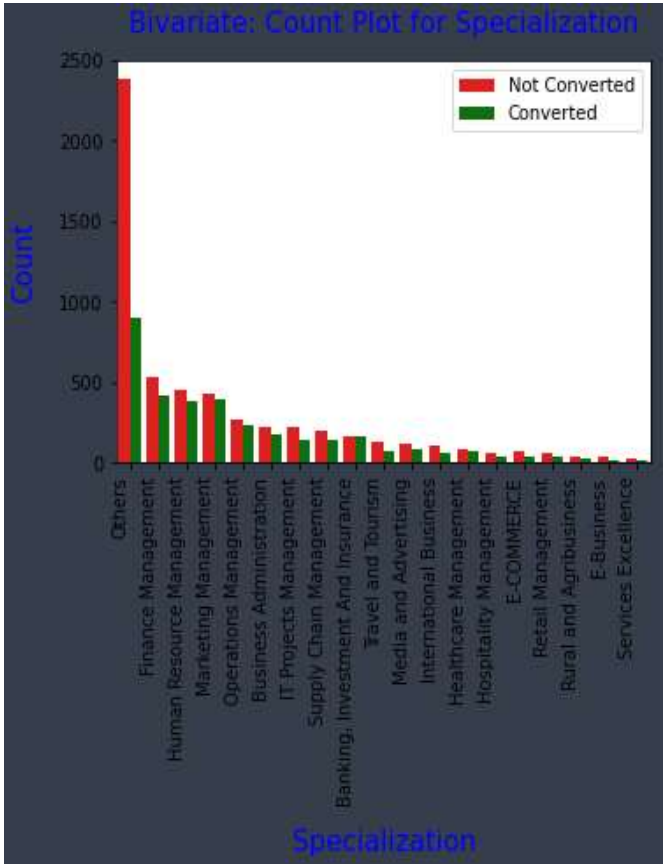




| Lead Origin | Converted |
|---|---|
| Lead Add Form | 93.63% |
| Landing Page Submission | 36.17% |
| API | 31.16% |
| Lead Import | 30.0% |

| Lead Source | Converted |
|---|---|
| Welingak Website | 98.45% |
| Reference | 92.55% |
| Google | 39.92% |
| Other Sources | 39.13% |
| Organic Search | 37.78% |
| Direct Traffic | 32.17% |
| Facebook | 29.03% |
| Olark Chat | 25.56% |
| Referral Sites | 24.8% |

# BIVARIATE ANALYSIS –CATEGORICAL COLUMNS

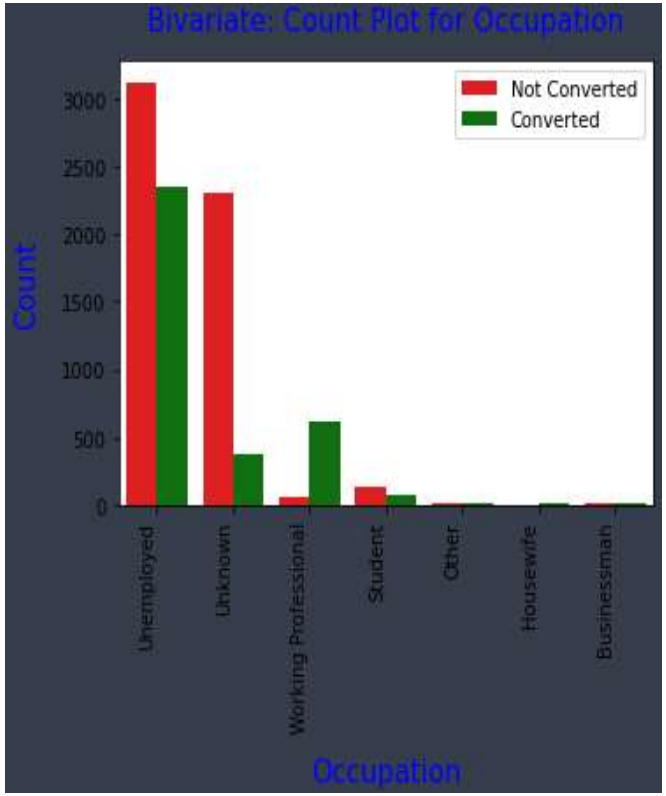
Most of the leads have not mentioned a specialization and around 27.45% of those converted
Leads with specialization in Banking, Investment And Insurance; Healthcare or Marketing - Over 48% Converted




| Specialization | Converted |
|---|---|
| Banking, Investment And Insurance | 48.96% |
| Healthcare Management | 48.72% |
| Marketing Management | 48.24% |
| Operations Management | 46.89% |
| Human Resource Management | 45.4% |
| Finance Management | 44.0% |
| Business Administration | 43.86% |
| Supply Chain Management | 42.77% |
| Rural and Agribusiness | 42.47% |
| Media and Advertising | 41.58% |
| Hospitality Management | 40.54% |
| IT Projects Management | 38.25% |
| E-Business | 36.84% |
| International Business | 35.23% |
| Travel and Tourism | 35.15% |
| E-COMMERCE | 35.14% |
| Retail Management | 34.0% |
| Services Excellence | 27.5% |
| Others | 27.45% |

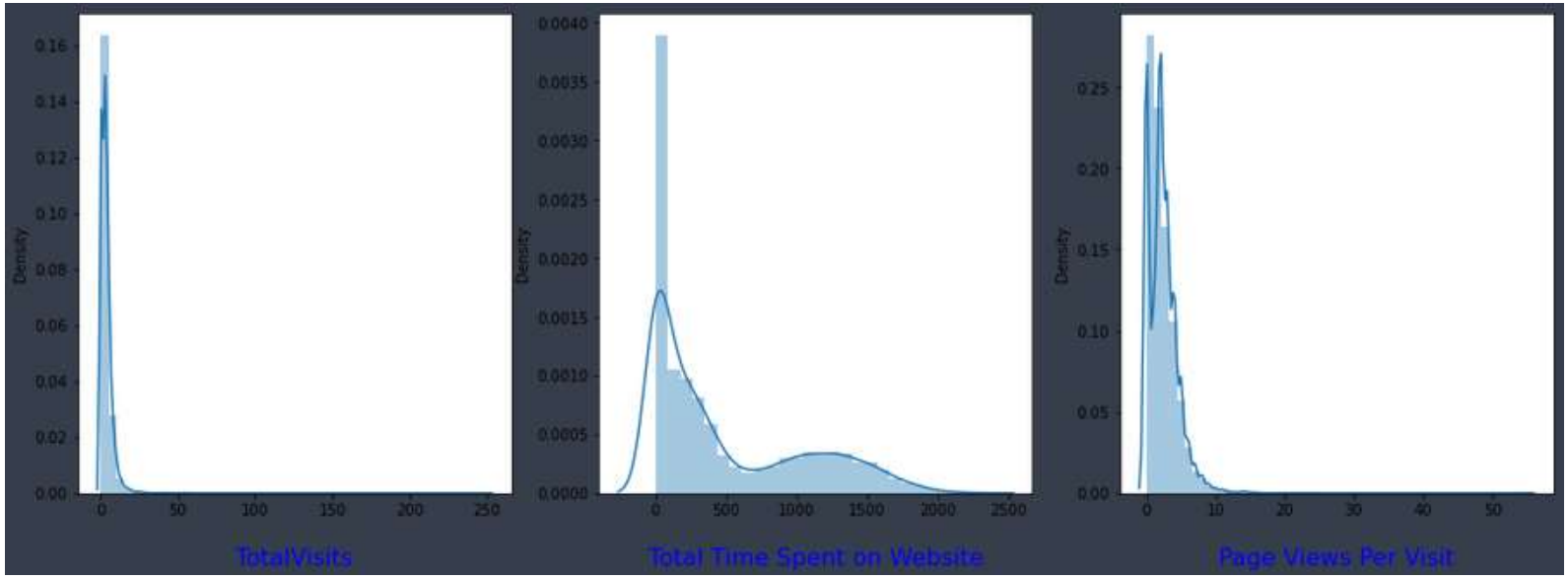
# BIVARIATE ANALYSIS –CATEGORICAL COLUMNS

Category Housewives are less in quantity, but have 100% conversion rate followed by Working professionals have >90% conversion rate
Though Unemployed people have been contacted in the highest number, the conversion rate is low (42.84%)




| Occupation | Converted |
|---|---|
| Housewife | 100.0% |
| Working Professional | 91.88% |
| Businessman | 62.5% |
| Other | 60.0% |
| Unemployed | 42.84% |
| Student | 35.92% |
| Unknown | 13.79% |

# UNIVARIATE ANALYSIS



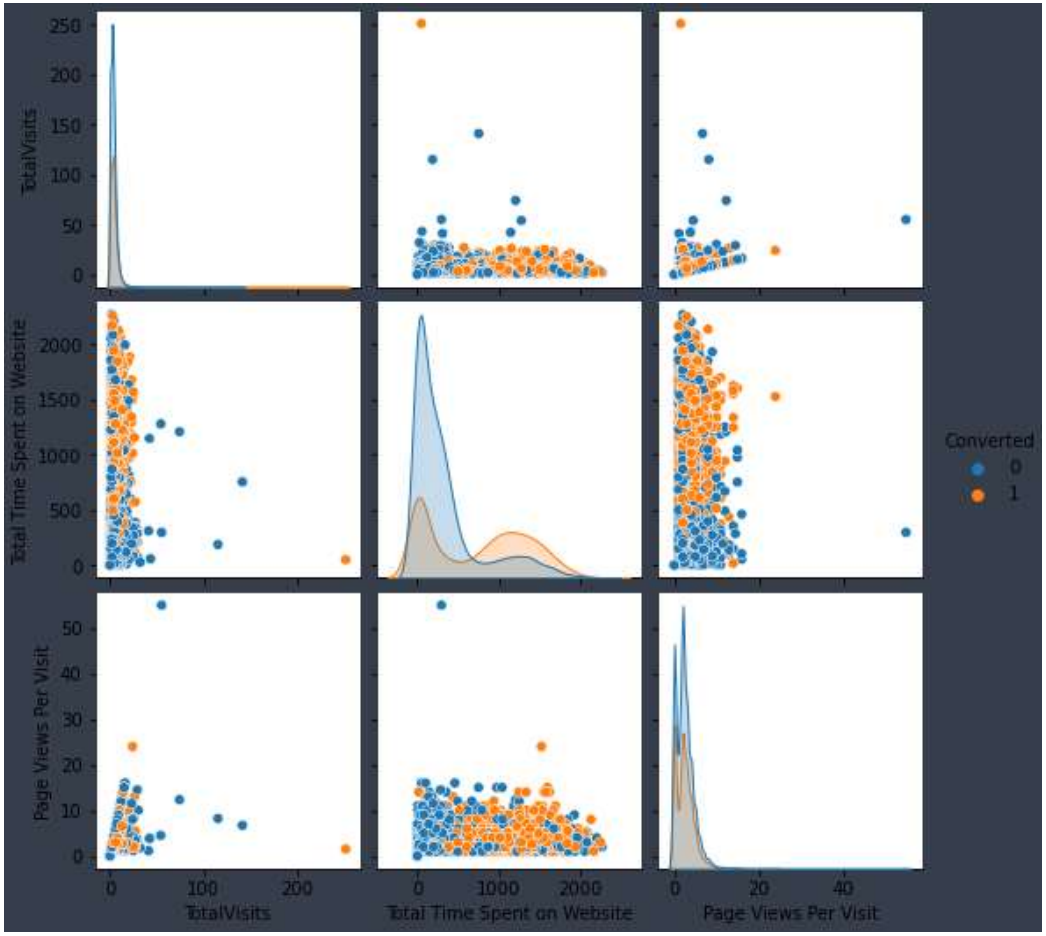

Data on Total Visits , Total Time Spent on Website and Page Views per Visit columns are not normally distributed & are significantly skewed.

Though outliers in TotalVisits and Page Views Per Visit shows valid values, this will misclassify the outcomes and consequently create problems when making inferences with the wrong model.

Logistic Regression is influenced by outliers. We will cap the TotalVisits and Page Views Per Visit to their 95th percentile due to 95th percentile and 99th percentile of these columns are very close and hence impact of the capping to 95th or 99th percentile will be the same.

# BIVARIATE ANALYSIS

Data is not normally distributed.
There are no linear relationship between the continuous features as shown by the above pair plot



# DATA CONVERSION

Converted “Do Not Email” binary variables (Yes/No) to (1/0)

```
No     8358
Yes     716
Name: Do Not Email, dtype: int64
```

```
0    8358
1     716
Name: Do Not Email, dtype: int64
```

Converted categorical variables with n levels to n-1dummy variables

```
Lead Origin                    = 4
Lead Source                    = 9
Specialization                 = 19
Occupation                     = 7
```

# MODEL BUILDING

- Splitting the Data into Training and Testing Sets we have chosen 70:30 ratio.
- Use RFE for Feature Selection with 15 variables as output
- Building Model by after checking p values (< 0.05) and VIF (< 5), the model is built multiple times to meet the criteria
- Predictions on test data set
- Checking the confusion matrix
- Overall accuracy 80.41%

```
[[3453  452]
 [ 792 1654]]
```

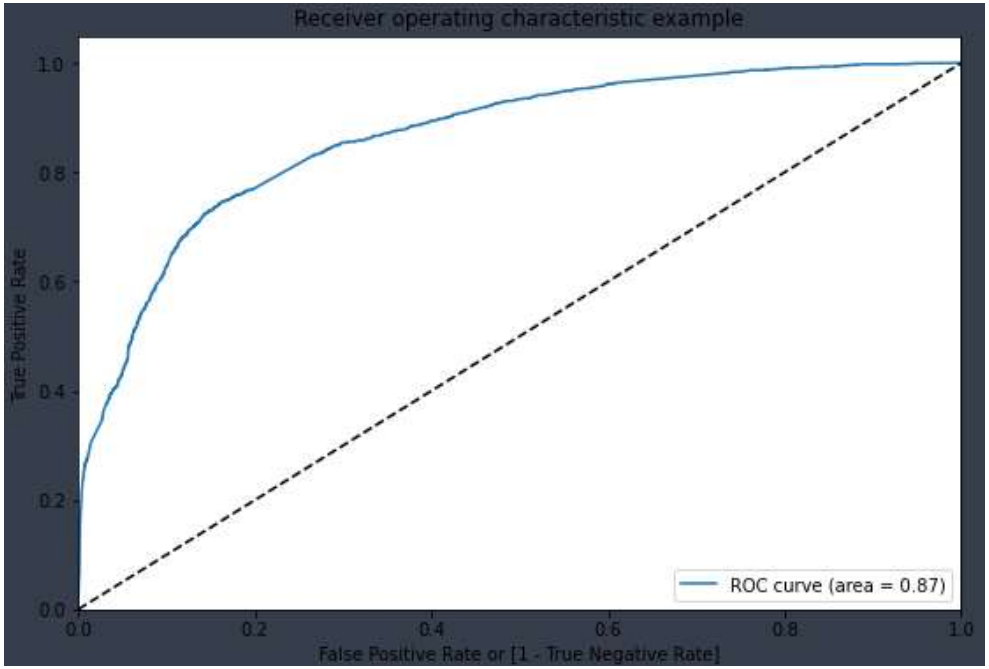
| | Features | VIF |
|---|---|---|
| 6 | Specialization_Others | 2.01 |
| 3 | Lead Source_Olark Chat | 1.90 |
| 7 | Occupation_Unknown | 1.58 |
| 2 | Lead Origin_Landing Page Submission | 1.37 |
| 1 | Total Time Spent on Website | 1.28 |
| 4 | Lead Source_Reference | 1.18 |
| 8 | Occupation_Working Professional | 1.18 |
| 0 | Do Not Email | 1.11 |
| 5 | Lead Source_Welingak Website | 1.09 |

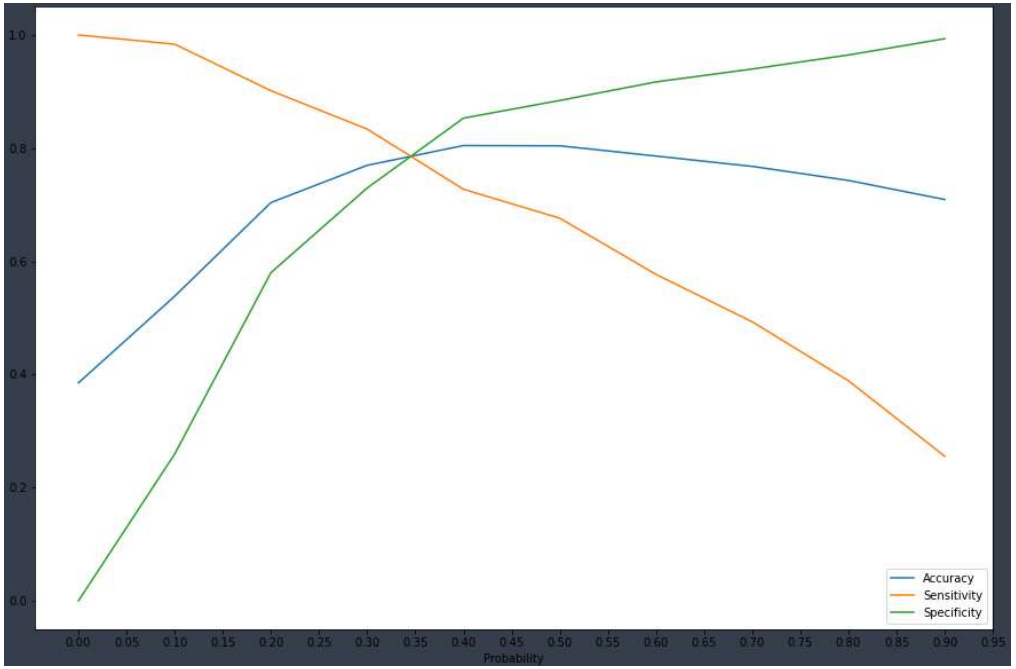
Generalized Linear Model Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | Converted | **No. Observations:** | 6351 |
| **Model:** | GLM | **Df Residuals:** | 6341 |
| **Model Family:** | Binomial | **Df Model:** | 9 |
| **Link Function:** | logit | **Scale:** | 1.0000 |
| **Method:** | IRLS | **Log-Likelihood:** | -2799.5 |
| **Date:** | Sun, 27 Jun 2021 | **Deviance:** | 5599.0 |
| **Time:** | 16:59:34 | **Pearson chi2:** | 6.43e+03 |
| **No. Iterations:** | 7 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | 0.0910 | 0.116 | 0.785 | 0.432 | -0.136 | 0.318 |
| **Do Not Email** | -1.4683 | 0.166 | -8.839 | 0.000 | -1.794 | -1.143 |
| **Total Time Spent on Website** | 1.1355 | 0.039 | 29.180 | 0.000 | 1.059 | 1.212 |
| **Lead Origin_Landing Page Submission** | -0.8487 | 0.121 | -7.002 | 0.000 | -1.086 | -0.611 |
| **Lead Source_Olark Chat** | 1.0723 | 0.115 | 9.295 | 0.000 | 0.846 | 1.298 |
| **Lead Source_Reference** | 3.4832 | 0.237 | 14.708 | 0.000 | 3.019 | 3.947 |
| **Lead Source_Welingak Website** | 5.9302 | 0.726 | 8.169 | 0.000 | 4.507 | 7.353 |
| **Specialization_Others** | -0.9541 | 0.119 | -8.043 | 0.000 | -1.187 | -0.722 |
| **Occupation_Unknown** | -1.3031 | 0.085 | -15.394 | 0.000 | -1.469 | -1.137 |
| **Occupation_Working Professional** | 2.3607 | 0.185 | 12.783 | 0.000 | 1.999 | 2.723 |

# ROC CURVE




The area under the Curve or Gini is 0.87 which is represent a good model
Optimal cut off probability is that probability where we get balanced sensitivity and specificity.
From the curve above, 0.35 is the optimum point to take it as a cutoff probability,
We will check the accuracy & confusion matrix after taking 0.35 as cutoff probability

# PRECISION AND RECALL TRADE-OFF

As per Precision-Recall Tradeoff, the cutoff is around 0.375 (between 0.35 and 0.40). We can choose the cut-off as 0.40 and use the Precision-Recall-Accuracy metrics to evaluate the model

# LEAD SCORE

Lead Score is calculated for all the leads in the Test data frame by the formula

Lead Score = Conversion Probability * 100

Higher the lead score, higher is the probability of a lead getting converted and vice versa

Since, we had used 0.40 as our final Probability threshold for deciding if a lead will convert or not, any lead with a lead score of 40 or above will have a value of '1' in the final_predicted column.

```
Y_pred_final['Lead_Score'] = Y_pred_final.Converted_prob.map( lambda x: round(x*100))
```

```
Y_pred_final.index = Y_pred_final.index.set_names(['Cust_Id'])
Y_pred_final.head()
```

| Cust_Id | Converted | Converted_prob | final_predicted | Lead_Score |
|---|---|---|---|---|
| 3271 | 0 | 0.054432 | 0 | 5 |
| 1490 | 1 | 0.978965 | 1 | 98 |
| 7936 | 0 | 0.045916 | 0 | 5 |
| 4216 | 1 | 0.927367 | 1 | 93 |
| 3830 | 0 | 0.060981 | 0 | 6 |

# MODEL EVALUATION STATISTICS

Comparison of Train & Test values

```
Train Data Accuracy    :80.02 %
Train Data Sensitivity :74.94 %
Train Data Specificity :83.2 %
Train Data F1 Score    :0.73
Test Data Accuracy     :80.46 %
Test Data Sensitivity  :74.94 %
Test Data Specificity  :83.2 %
Test Data F1 Score     :0.73
```

Classification Report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.84 | 0.83 | 0.84 | 3905 |
| 1 | 0.74 | 0.75 | 0.74 | 2446 |
| accuracy | | | 0.80 | 6351 |
| macro avg | 0.79 | 0.79 | 0.79 | 6351 |
| weighted avg | 0.80 | 0.80 | 0.80 | 6351 |

# CONCLUSION

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

- Increase user engagement with users receiving Emails
- Get Total Time Spent on Website increased by advertising and user experience which makes the customer engaging in the website as it contributes in higher conversion
- Improve the Welingark Website since this is affecting the conversion negatively
- Focus less on Hospitality Management Specialisation as it is majorily a physical service based industry ex Front desk, Cooking etc & cannot be learnt simply over an online course