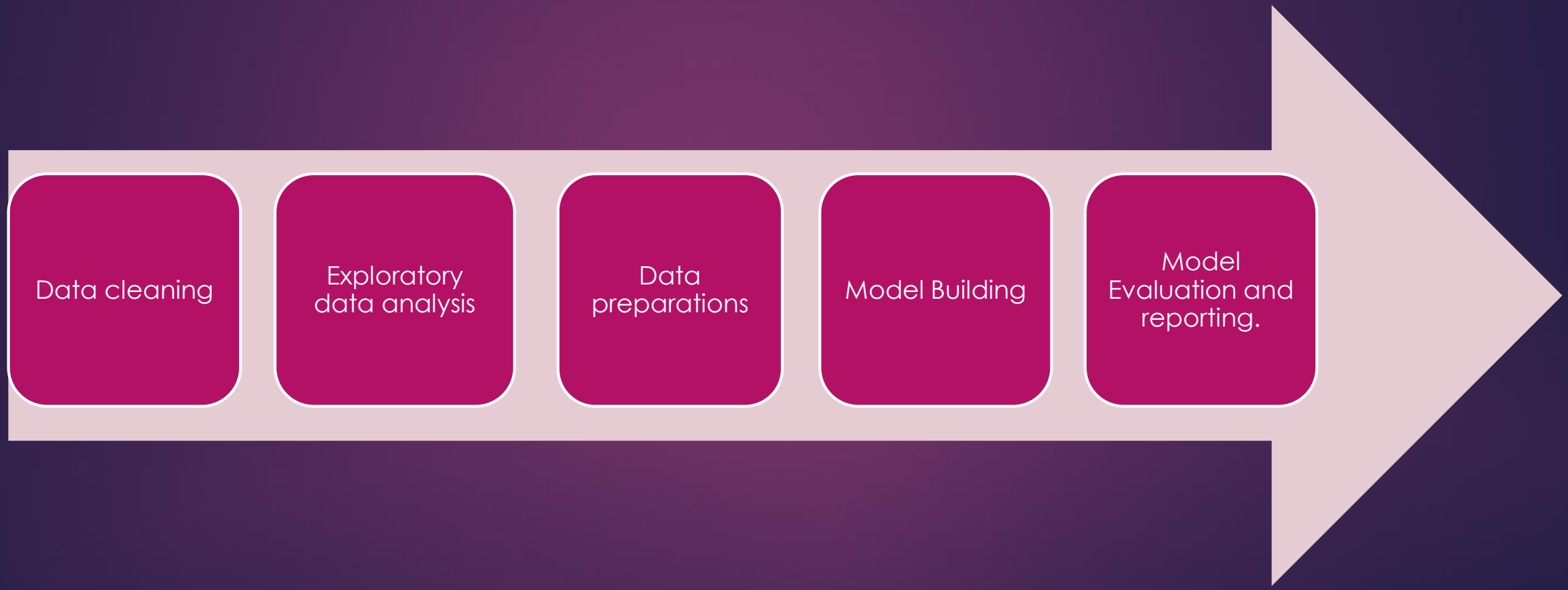
# Lead score case study

BY:- APARNESH MURTHY & SHASHANK SEKHAR

# Business Objectives

- X education company is experiencing a drop in the lead conversion rate.
- They want to identify the leads with the most potential of conversion in order to make their sales process more efficient.
- The target conversion rate for the company is around 80 %. A logistic regression model is needed in order to assign each lead a score between 0 -100 and identifying the best ones on its basis.

# Solution Methodology



# Data Cleaning

## Identifying score variables and dropping them:-

- These variables are generated by the sales team after discussion with a lead
- Examples:- Lead quality , Asymmetrique Activity Index, Asymmetrique Profile Index etc.

## Identifying columns with missing data :-

- Based on the %age of missing data columns are either removed or imputed with suitable values.
- Columns having missing data greater than 40 % were dropped.

## Identifying columns having highly skewed data :-

- Categorical columns which contain an extremely high percentage of occurrence of certain category are dropped.
- Examples of such columns are:- Search, Magazine, Newspaper articles etc.

# Data Cleaning continuing

## Identifying variables where values can be clubbed:-

These are categorical variables which are distributed well. But they have a few categories which can be clubbed as they have fewer occurences.

Examples:- Last activity, Lead Source, Specialisation etc.

## Identifying rows with value "Select" and treating them as missing data :-

Some rows have a value "select" which means no entry was added to it. For rows such as these they would be replaced with NaN value.

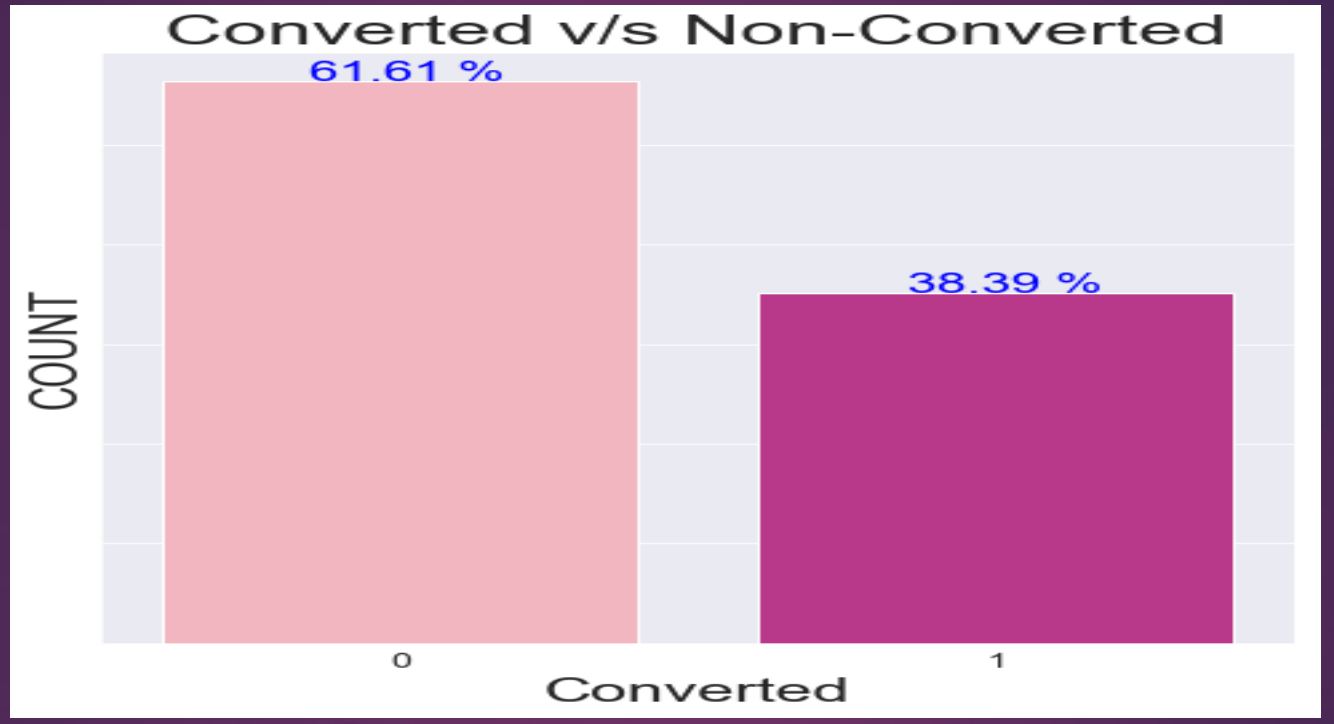
These rows will be treated as missing data and will be taken into account when treating columns with missing data.

## Treating variables with outliers:-

Variables with outliers are treated by dropping the values beyond the 99 % quantile.

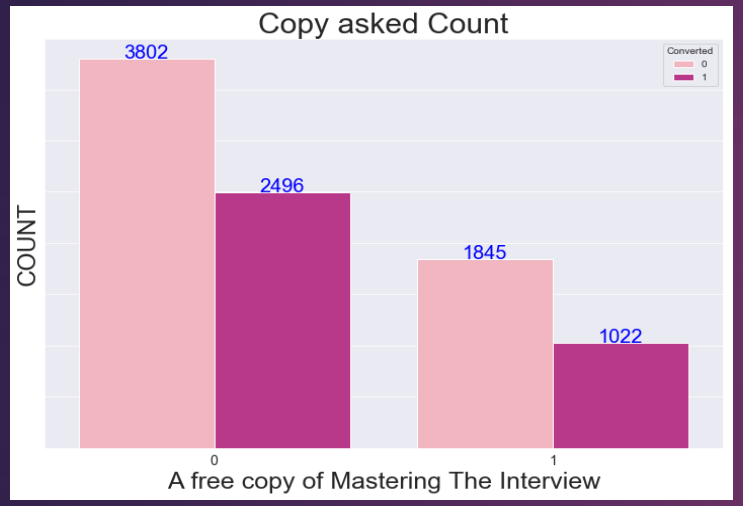
Examples of such columns are:- Total visits, Page Views per visit etc.

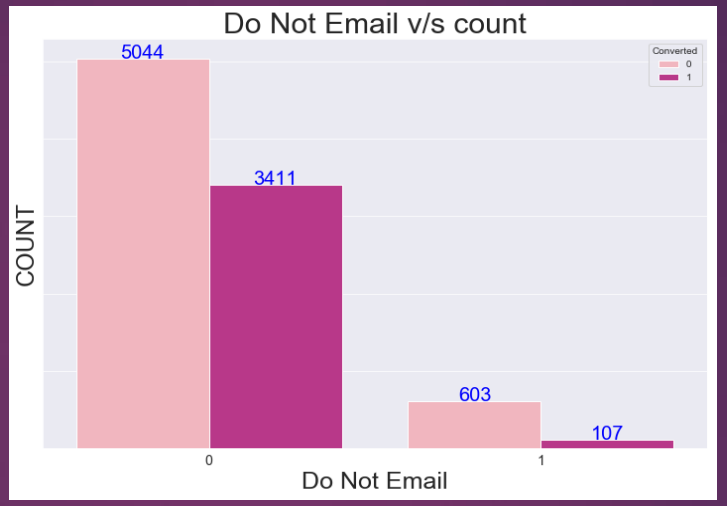
## Exploratory data analysis

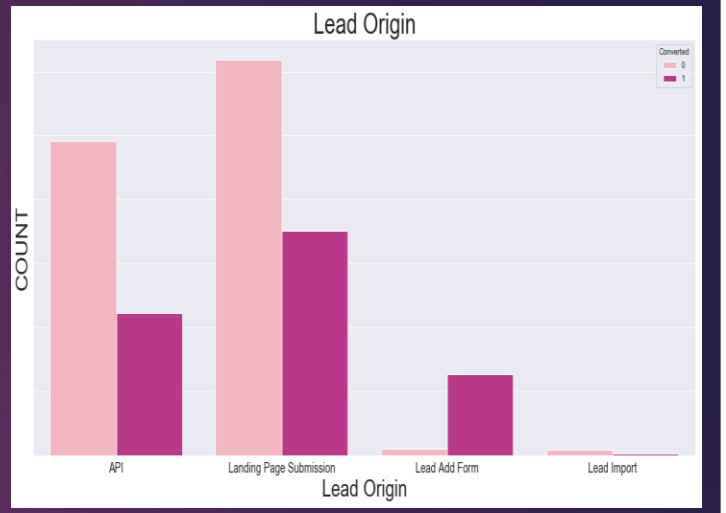



Current lead conversion ratio

# Categorical variable analysis

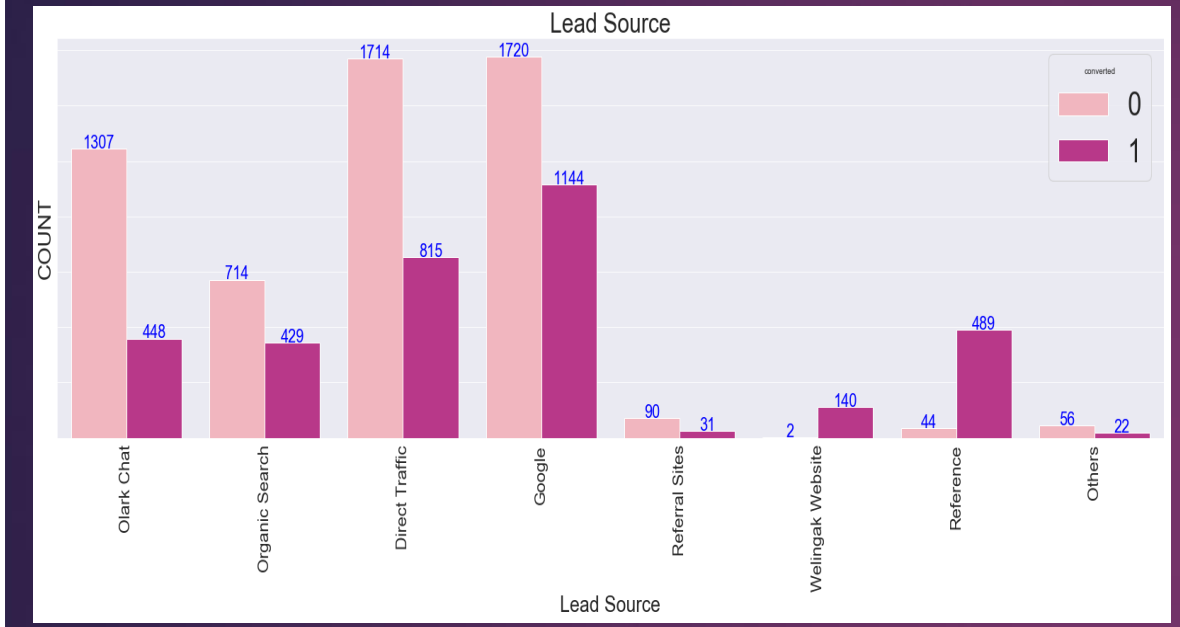


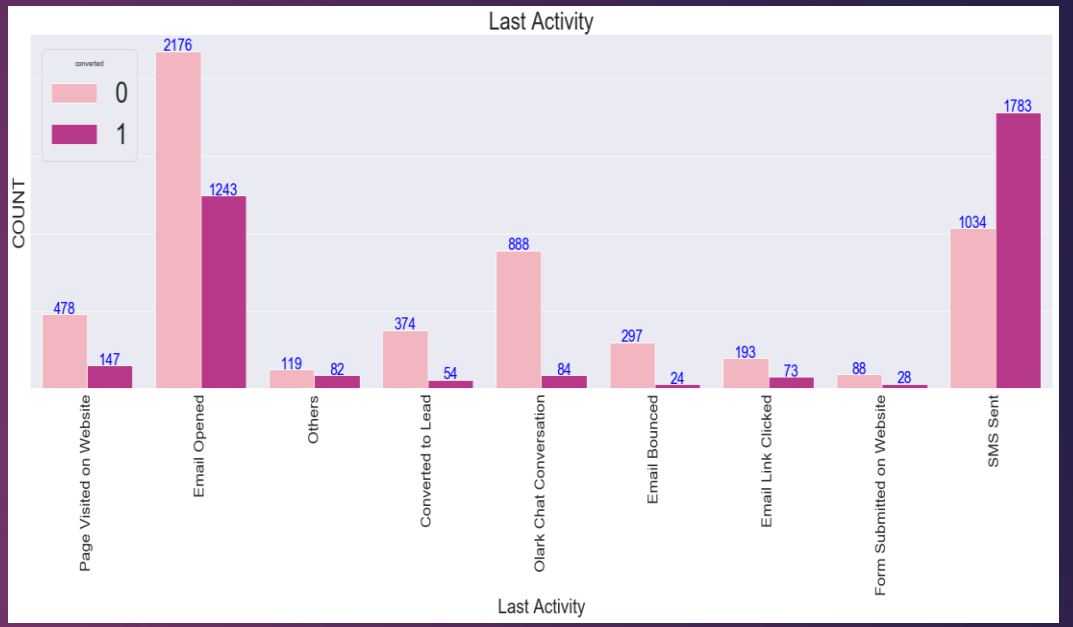




- People who don't want a copy are more likely to opt for the course.
- A call is more likely to be successful if the user has opted for the email service.
- Most leads are identified by the landing page submission, but the highest conversion rate is of the leads which were identified by lead add form.

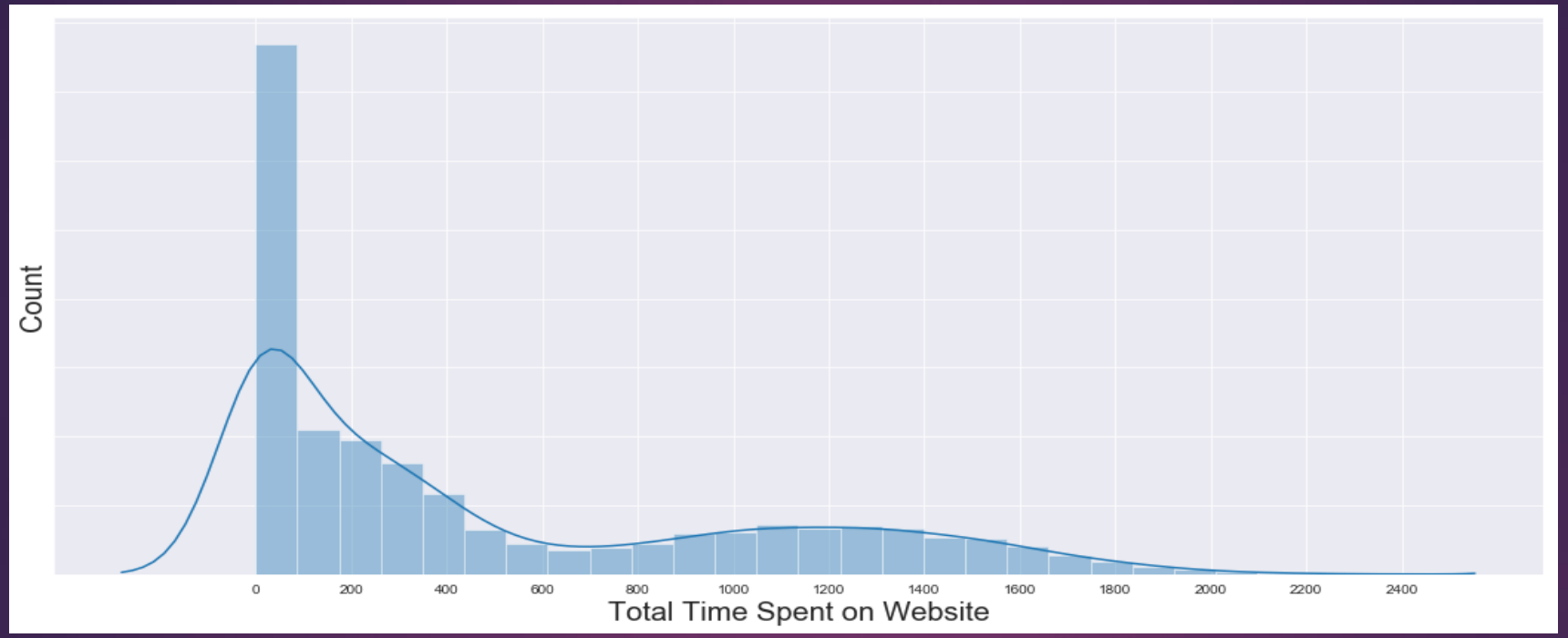
# Categorical variable analysis continuing



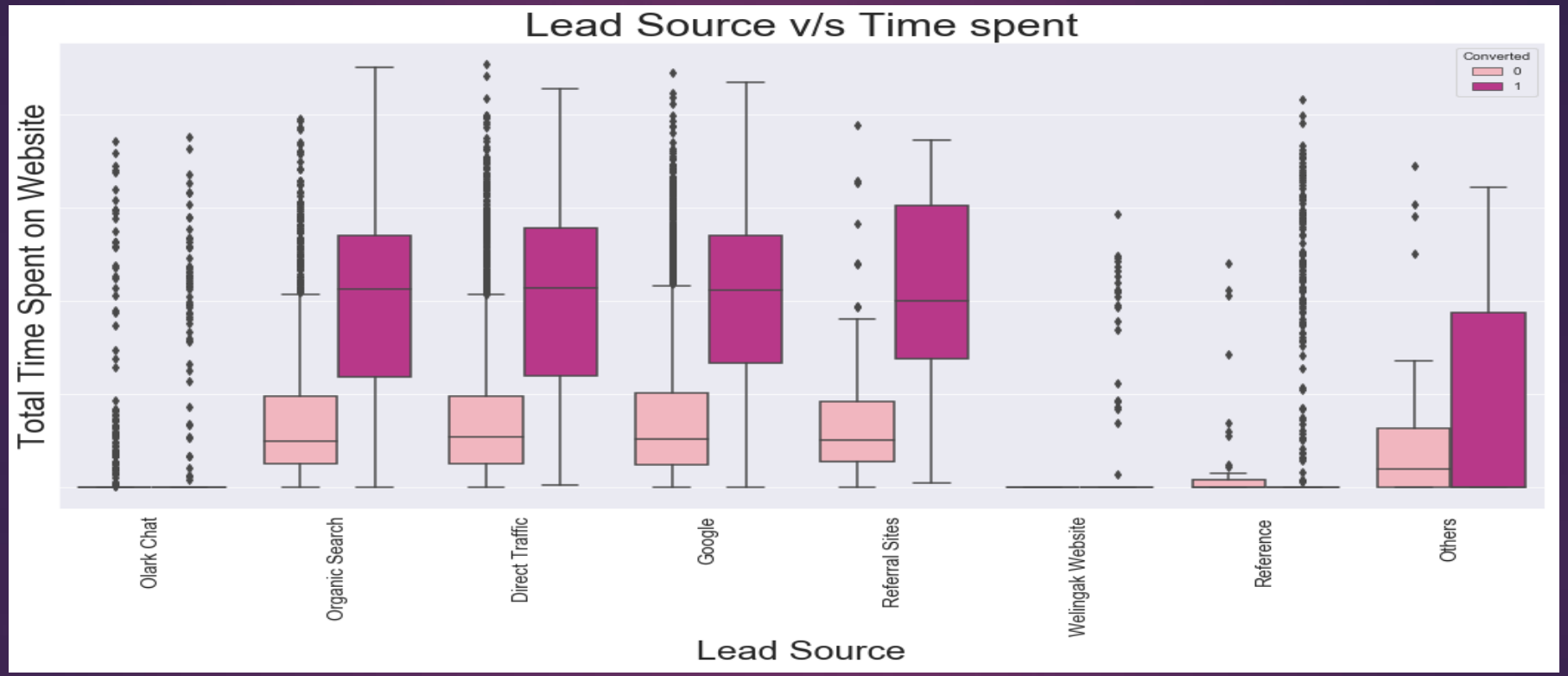

- People who learned about the course through references have the highest conversion rate.
- Conversion rate is the highest of leads who get communications regarding the courses through SMS.

# Numerical variable analysis



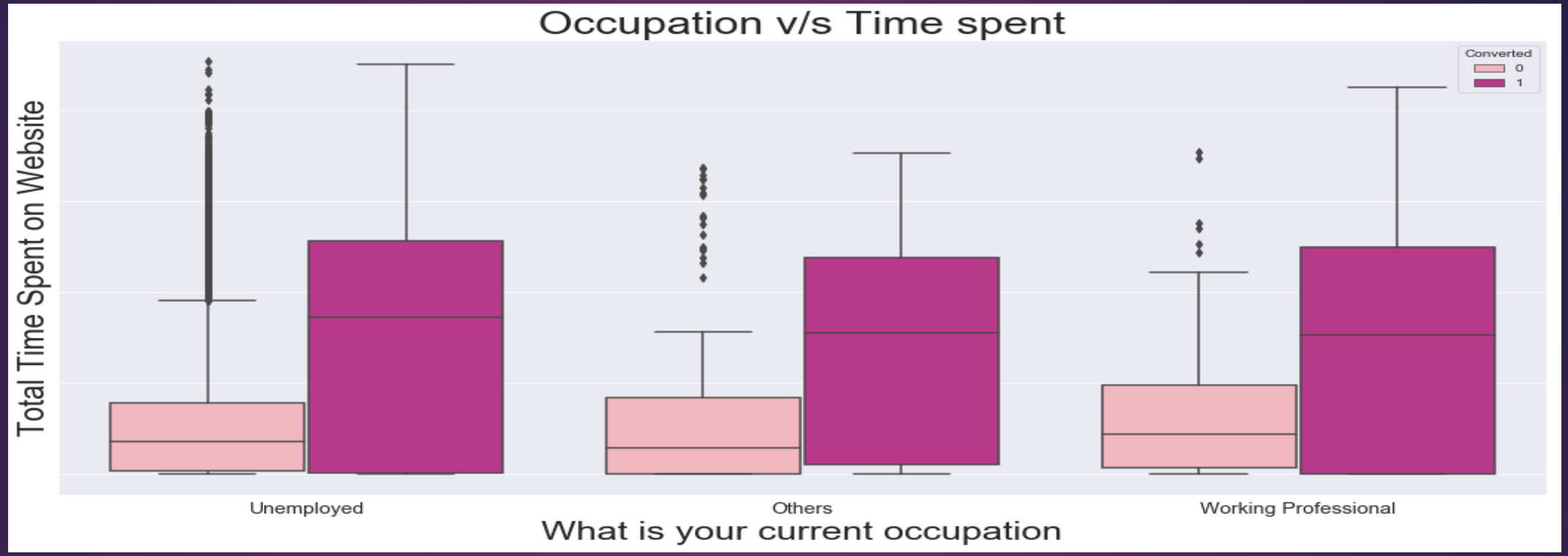

- Most of the users spend less than 250 minutes on the website

# Bi - variate analysis on categorical and numerical variables



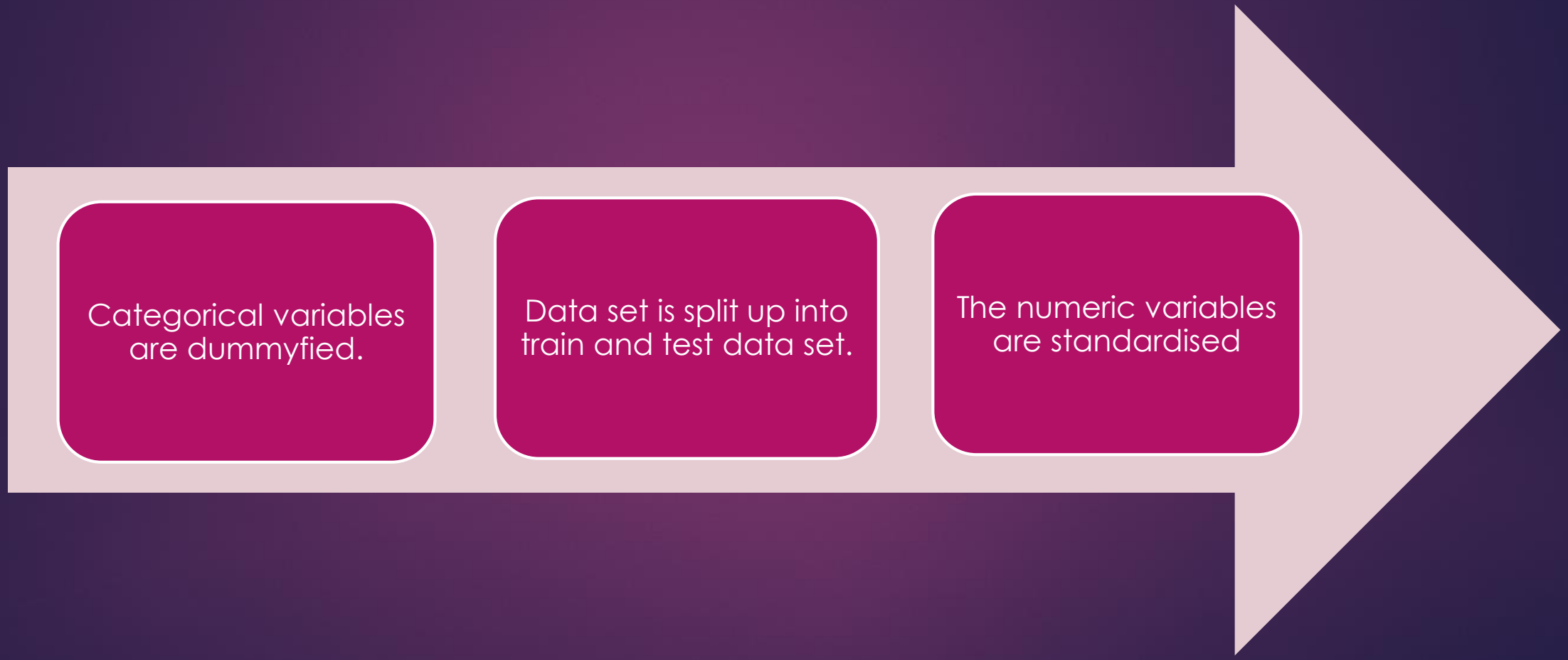

- People who landed on the website via Organic search,google, referral sites and spend significantly more amount of time are most likely to opt for the course

## Bi - variate analysis on categorical and numerical variables




- People who spend a good amount of time on the website are more likely to get converted into a student irrespective of the occupational status.

# Data Preparation



# Model building

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -1.9412 | 0.156 | -12.433 | 0.000 | -2.247 | -1.635 |
| Do Not Email | -1.7020 | 0.195 | -8.738 | 0.000 | -2.084 | -1.320 |
| TotalVisits | 1.2209 | 0.309 | 3.948 | 0.000 | 0.615 | 1.827 |
| Total Time Spent on Website | 4.5253 | 0.165 | 27.459 | 0.000 | 4.202 | 4.848 |
| Lead Origin_Landing Page Submission | -1.0371 | 0.125 | -8.279 | 0.000 | -1.283 | -0.792 |
| Lead Origin_Lead Add Form | 3.4319 | 0.220 | 15.606 | 0.000 | 3.001 | 3.863 |
| Lead Source_Olark Chat | 1.2331 | 0.129 | 9.565 | 0.000 | 0.980 | 1.486 |
| Lead Source_Welingak Website | 2.5664 | 0.755 | 3.399 | 0.001 | 1.086 | 4.047 |
| Last Activity_Email Opened | 0.4850 | 0.092 | 5.289 | 0.000 | 0.305 | 0.665 |
| Last Activity_Olark Chat Conversation | -1.0623 | 0.179 | -5.933 | 0.000 | -1.413 | -0.711 |
| Last Activity_Others | 0.5703 | 0.290 | 1.967 | 0.049 | 0.002 | 1.138 |
| Specialization_Not Available | -1.0805 | 0.122 | -8.870 | 0.000 | -1.319 | -0.842 |
| What is your current occupation_Working Professional | 2.5787 | 0.189 | 13.615 | 0.000 | 2.208 | 2.950 |
| Last Notable Activity_Others | 1.4786 | 0.345 | 4.291 | 0.000 | 0.803 | 2.154 |
| Last Notable Activity_SMS Sent | 2.0195 | 0.101 | 20.036 | 0.000 | 1.822 | 2.217 |

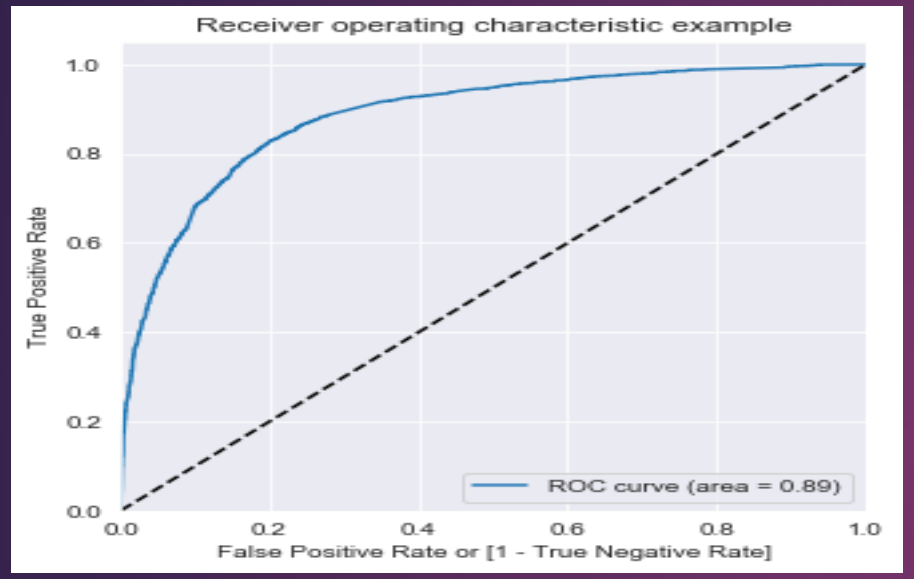
| | Column | VIF |
|---|---|---|
| 3 | Lead Origin_Landing Page Submission | 3.24 |
| 1 | TotalVisits | 2.63 |
| 10 | Specialization_Not Available | 2.46 |
| 7 | Last Activity_Email Opened | 2.22 |
| 2 | Total Time Spent on Website | 2.14 |
| 5 | Lead Source_Olark Chat | 2.14 |
| 13 | Last Notable Activity_SMS Sent | 1.81 |
| 8 | Last Activity_Olark Chat Conversation | 1.67 |
| 12 | Last Notable Activity_Others | 1.53 |
| 4 | Lead Origin_Lead Add Form | 1.44 |
| 9 | Last Activity_Others | 1.44 |
| 0 | Do Not Email | 1.29 |
| 6 | Lead Source_Welingak Website | 1.26 |
| 11 | What is your current occupation_Working Profes... | 1.20 |

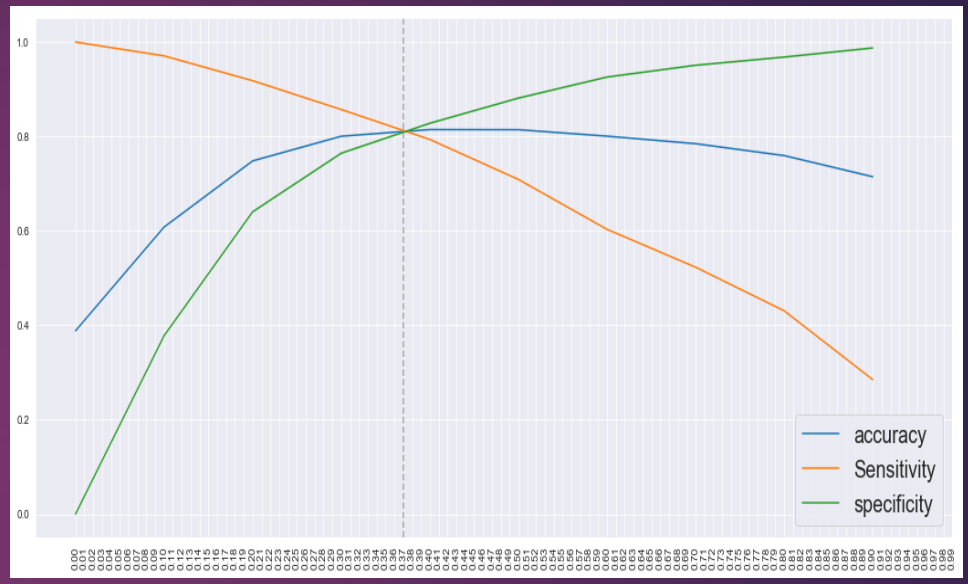
- The initial 20 variables for the logistic regression model are chosen using RFE.
- The cut-off p-values for the test of significance of the variables in the model is chosen at 0.05 and the cut-off for VIF to test for Multicollinearity between independent variables is chosen at 5. The final model satisfies the cut-off for all these tests.

# Model building


- The correlation heat map for the final model with 16 variables confirms that the independent variables are not correlated

# Model building




- The ROC curve seems to have a good AUC. Hence ,we can say that the model seems to be a good fit.
- The optimal cut-off point for the model is chosen on the basis of the evaluation metric which in the case of this case study is :- sensitivity. As, the main goal of creating this model is to increase the incisiveness by which it can identify the leads who will convert. So, prioritising sensitivity seems to be the best course of action.
- The graph on the right suggests that p=0.37 might be a good cut-off as the sensitivity and accuracy on the train set comes up to be 0.8 and 0.79.

# Model building

| | prob | accuracy | sensitivity | specificity |
|---|---|---|---|---|
| **0.0** | 0.0 | 0.388309 | 1.000000 | 0.000000 |
| **0.1** | 0.1 | 0.608106 | 0.970695 | 0.377931 |
| **0.2** | 0.2 | 0.748246 | 0.918105 | 0.640418 |
| **0.3** | 0.3 | 0.800468 | 0.857086 | 0.764526 |
| **0.4** | 0.4 | 0.814653 | 0.793657 | 0.827982 |
| **0.5** | 0.5 | 0.814341 | 0.708952 | 0.881244 |
| **0.6** | 0.6 | 0.800624 | 0.603372 | 0.925841 |
| **0.7** | 0.7 | 0.784723 | 0.523083 | 0.950815 |
| **0.8** | 0.8 | 0.759470 | 0.430751 | 0.968145 |
| **0.9** | 0.9 | 0.714731 | 0.285026 | 0.987513 |

- However, as mentioned earlier, maximising the sensitivity while maintaining a reasonable score of accuracy and other evaluation metrics is the necessity for solving this business problem.
- The following table shows the value of accuracy, sensitivity and specificity of the model at different cut-offs of probability on the training set. We can see that at p=0.3 we are getting a great value of sensitivity while also maintaining its accuracy and specificity.
- The sensitivity and accuracy at the cut-off p= 0.3 on the training set is 0.83 and 0.8 respectively which is better than what we got at p=0.37 from the graph previously. So we are choosing this for our predictions.

## Evaluating the model on train dataset

| # Predicted<br># Actual | Not converted | Converted |
|---|---|---|
| Not Converted | 3000 | 924 |
| Converted | 356 | 2135 |

Probability threshold :- 0.3

| Accuracy | Sensitivity | Specificity | FPR |
|---|---|---|---|
| 0.8 | 0.85 | 0.76 | 0.23 |

| TPR | Precision | Recall |
|---|---|---|
| 0.70 | 0.70 | 0.85 |

## Evaluating the model on test dataset

| # Predicted<br># Actual | Not converted | Converted |
|---|---|---|
| Not Converted | 3000 | 924 |
| Converted | 356 | 2135 |

Probability threshold :- 0.3

| Accuracy | Sensitivity | Specificity | FPR |
|---|---|---|---|
| 0.8 | 0.83 | 0.78 | 0.22 |

| TPR | Precision | Recall |
|---|---|---|
| 0.69 | 0.69 | 0.83 |

# Reporting final lead score calculated on the test set

| | prospectID | Converted | Converted_prob | final_predicted | Lead_score |
|---|---|---|---|---|---|
| **0** | 4811 | 0 | 0.278245 | 0 | 27.82 |
| **1** | 2606 | 0 | 0.606297 | 1 | 60.63 |
| **2** | 8394 | 1 | 0.585334 | 1 | 58.53 |
| **3** | 7068 | 0 | 0.319183 | 0 | 31.92 |
| **4** | 7695 | 0 | 0.744407 | 1 | 74.44 |

- The lead score as calculated on the test set and the predictions made on it based on the cut-off value of p=0.3.