EXECUTIVE PG PROGRAMME IN
DATA SCIENCE (IIIT-B)

# Lead Scoring Case Study

*By : Derin David C and Shreyas Dubey*
*Batch : DS_C32*

## Problem

An education company named X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. Once these people land 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. 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 so they want to improve it.

## Objective

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 and the customers with lower lead score have a lower conversion chance.

# Steps Involved

• Importing the data : In this step we import all the necessary libraries and the converted CSV file of the data into a data frame.

• Inspecting the data : In this step we inspected the data for its shape, missing values, outliers and data types of the columns.

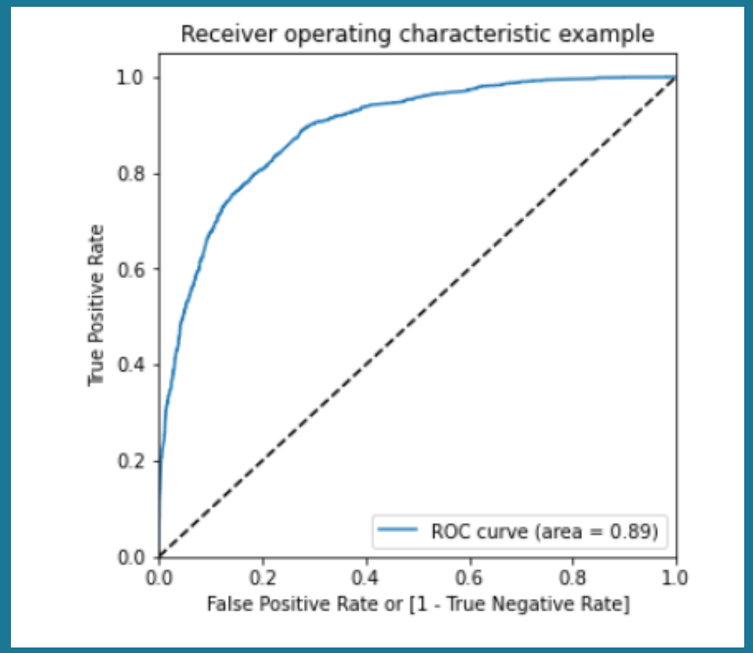
• Data Preparation : In this step we handled missing values in all the columns using various techniques like median or mode replacements and making a new category of the missing variables and in some case dropping the column if the missing values are higher than 40%. As there were no visible outliers in the data we concluded that the data is in proper form for applying EDA.

• EDA and Data Modifications : In this step we applied EDA on the given data and then modify the data based on our analysis such as reducing the categories In a column or dropping a column etc. We Also came out with some amazing insights from the data.

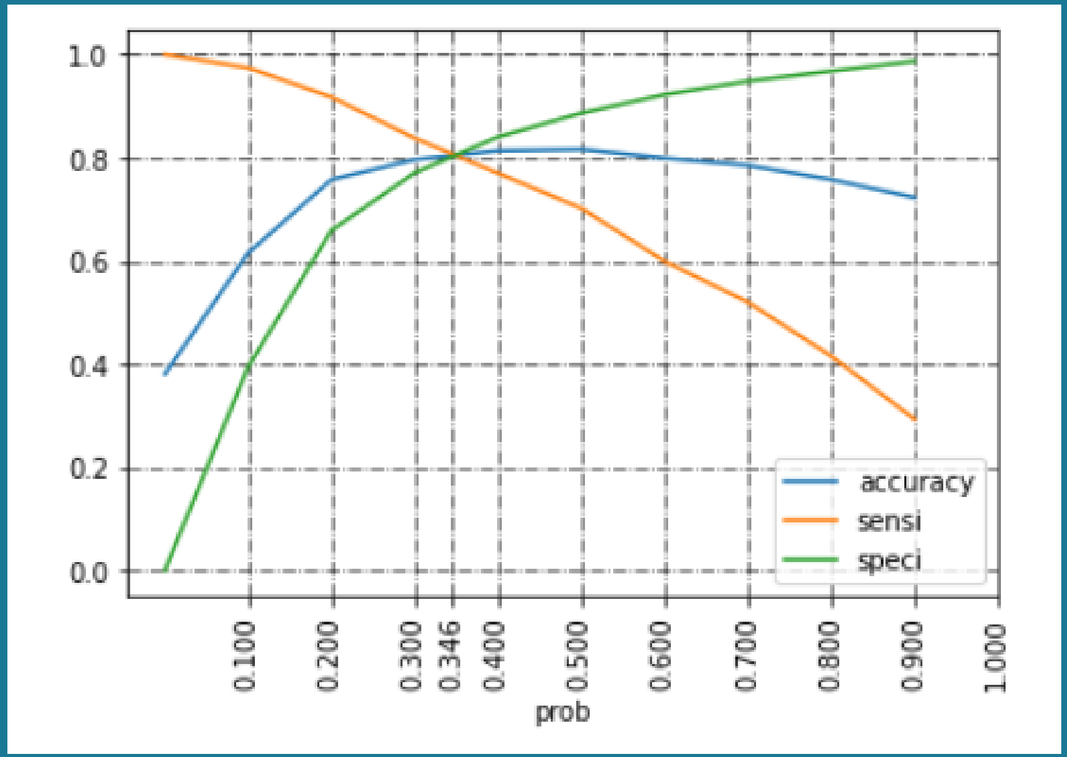
• Preparing data for modeling : In this step we will prepared the data for modeling, which includes Dummy variable creation for categorical variables, Test Train Split (30-70 split )and Standard Scaling for  numerical variables.

• Model building : In this step we choose the top 15 variables using Recursive feature elimination technique(RFE) and them reduced the feature using manual approach using Logistic model from stats module with filter such as P-values < 0.05 and VIF < 5.

• Making ROC curve : After making model we checked the model using ROC curve in which we got an area under the curve of (AUC) = 0.89 which is good for a logistic regression model.



• Finding Optimal Cutoff Point : In this step we find the cut-off point for the model by plotting accuracy, sensitivity and specificity for various cut-offs. We concluded that 0.3 is a good cut-off for our analysis.We finally got a sensitivity of 0.84 for the training set. ( In this case we take sensitivity as our metric as we want to reduce the false negative values).



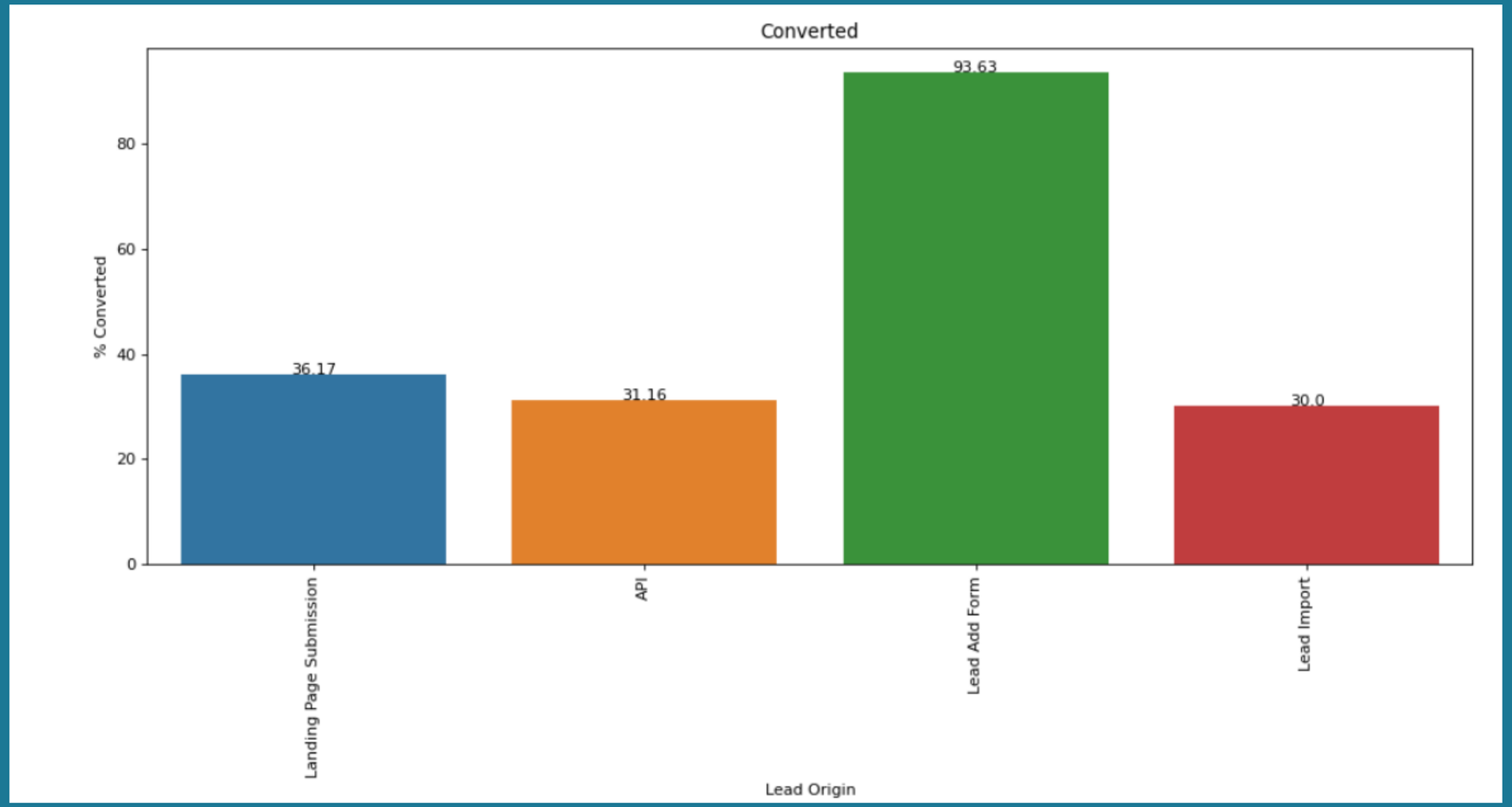
- Making predictions on the test set : Finally we predict using the final model on the test set and achive sensitivity score of 0.83 on the test set.

Final Scores

| Metrics | Accuracy |
|---|---|
| Train_Accuracy_Score | 0.80 |
| Train_Sensititvity_Score | 0.84 |
| Train_Specificity_Score | 0.77 |
| Test_Accuracy_Score | 0.80 |
| Test_Sensititvity_Score | 0.83 |
| Test_Specificity_Score | 0.78 |

# Analysing Each Column

# Analysis of Lead Origin

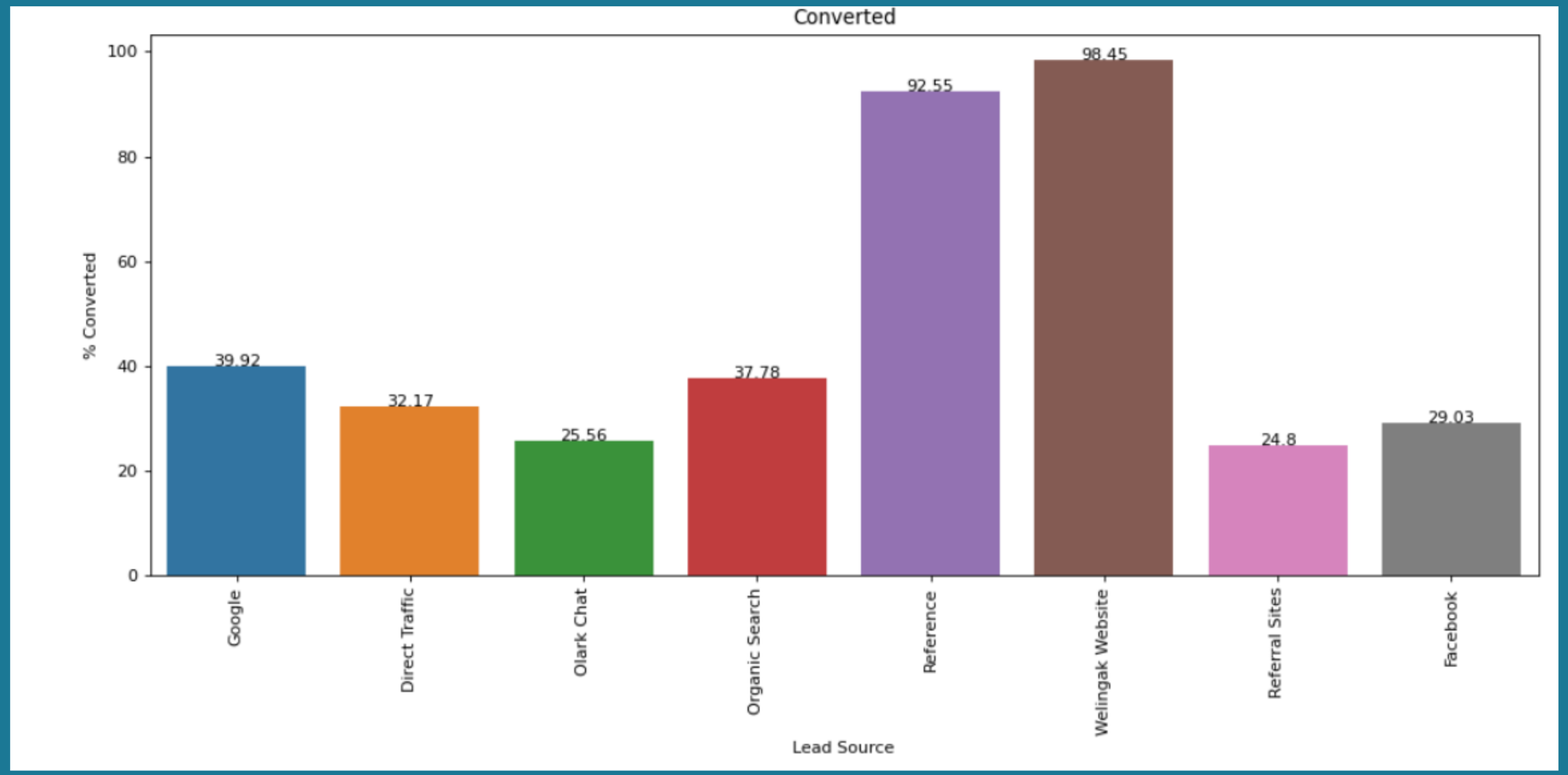



Inference :

Lead Add Form category have highest % of convertions.

Landing Page Submission category gives highest no of conventions

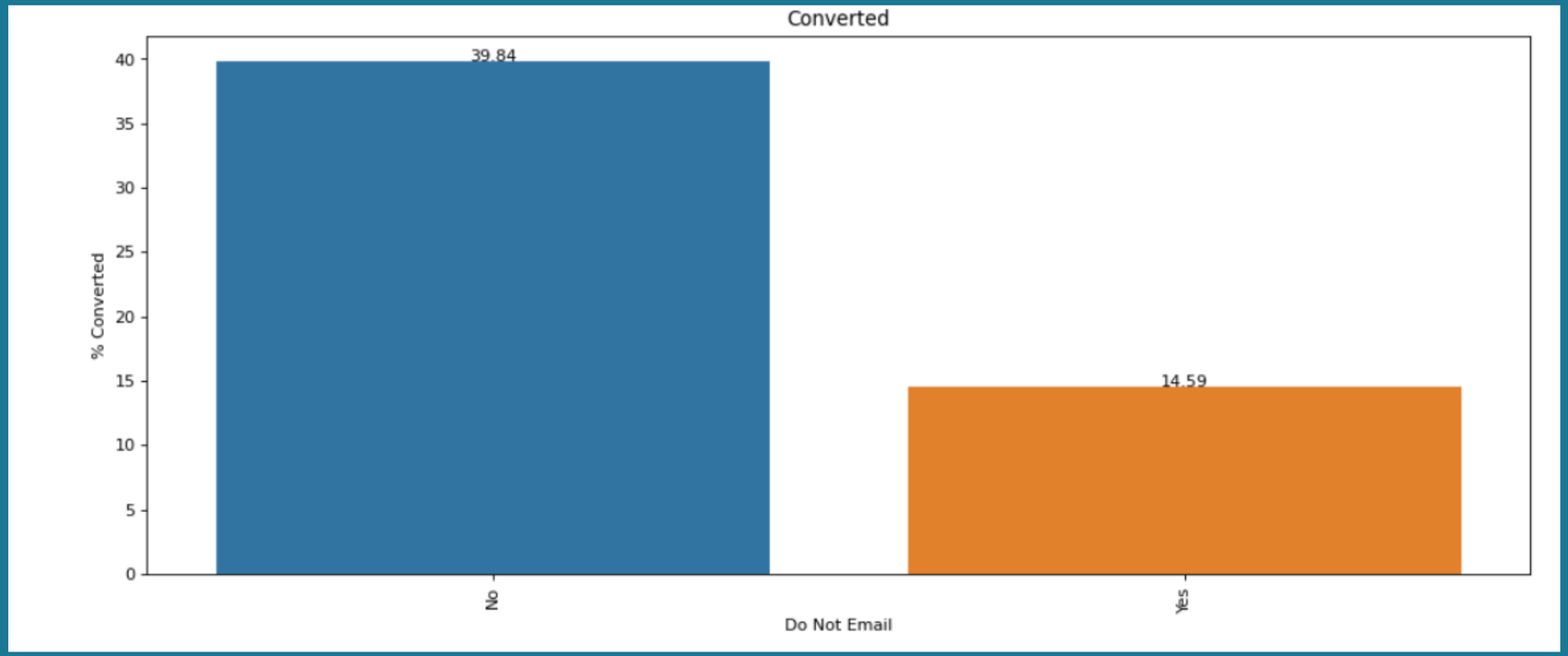
# Analysis of LeadSource




Inference :

- Highest % of convertion are from Reference and Welingak Website category.
- Highest no of convertions are from Google and Direct Traffic category.

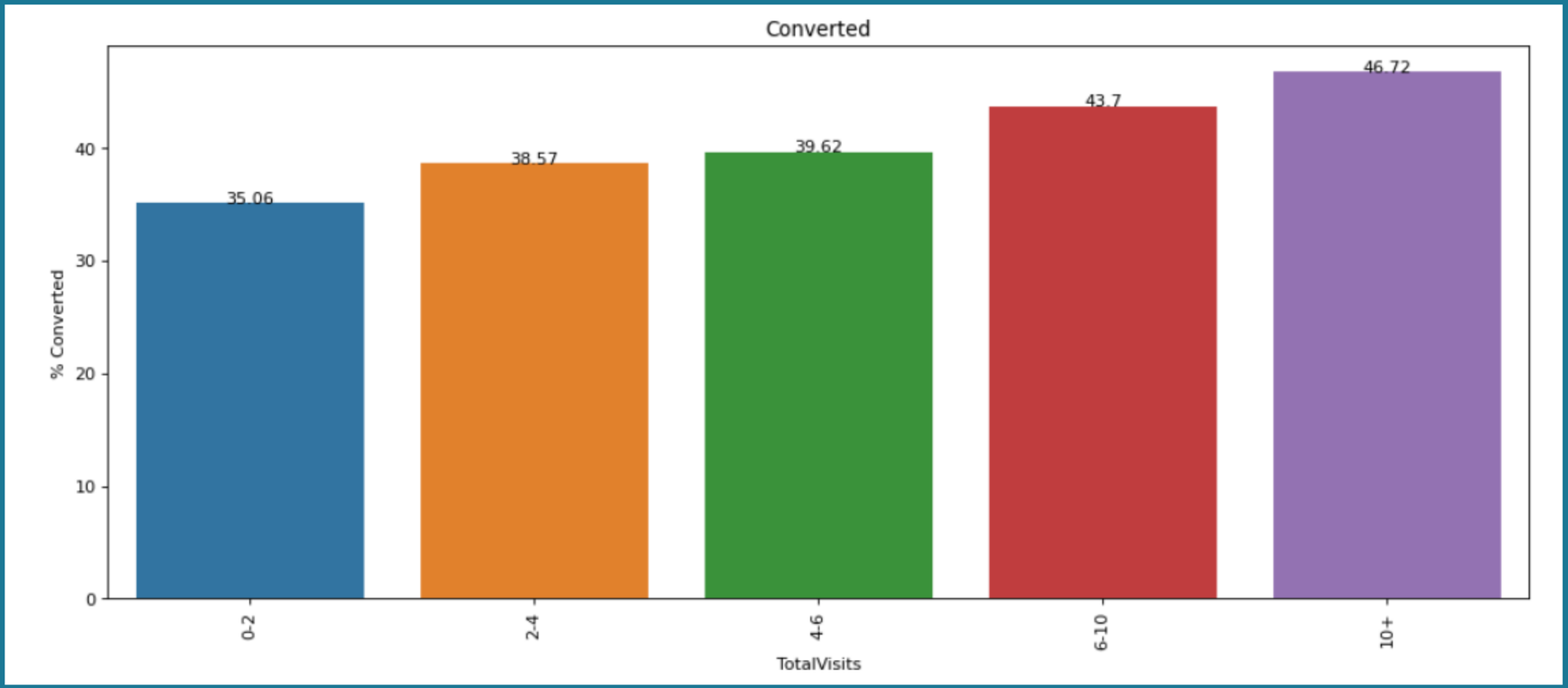
# Analysis of Do Not Email




Inference :

- people who say no to Do not email have clearly high chance to convert.

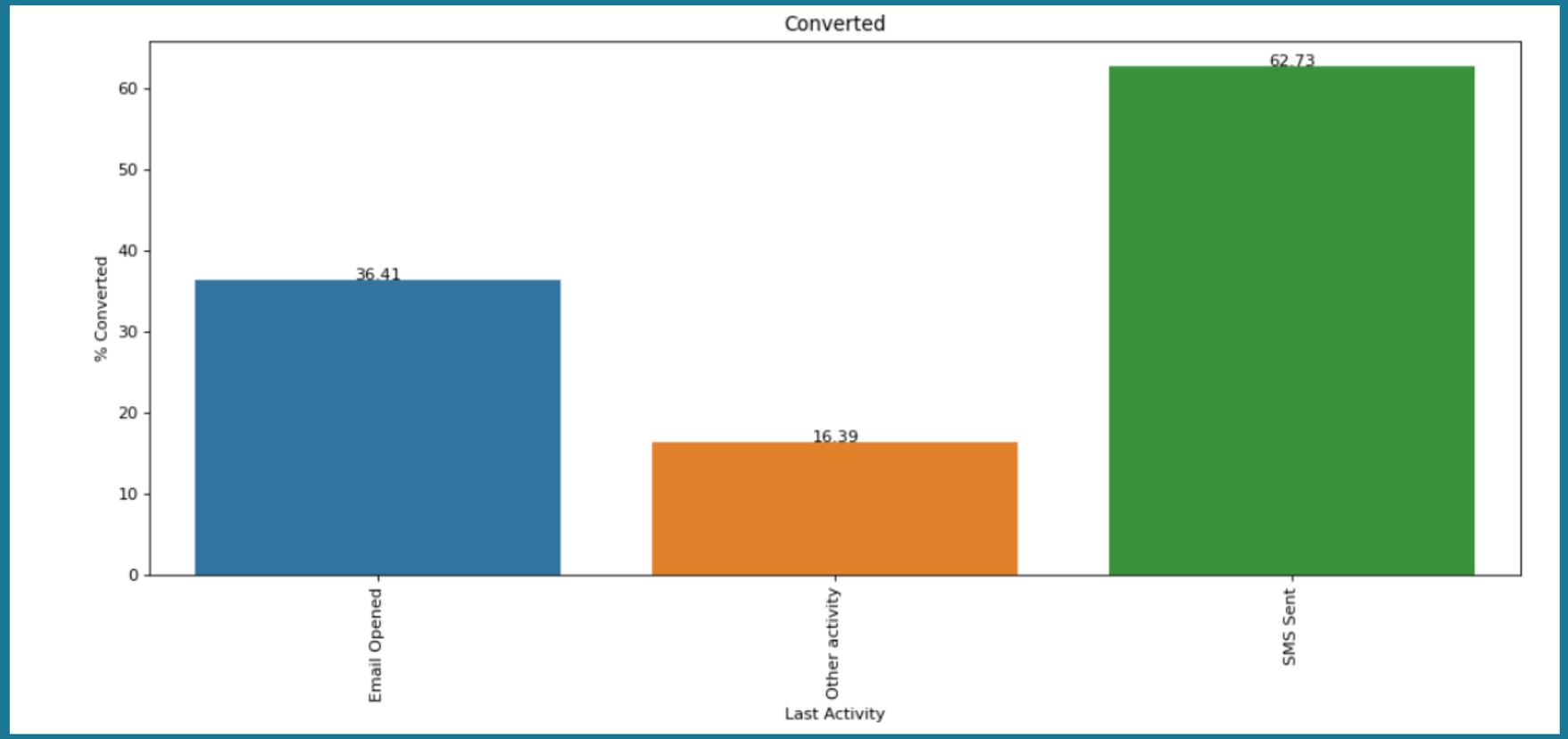
# Analysis of TotalVisits




Inference :

- probablity of convertions increases with the increase in total visits.
- We are getting highest no of convertion in 2-4 category.

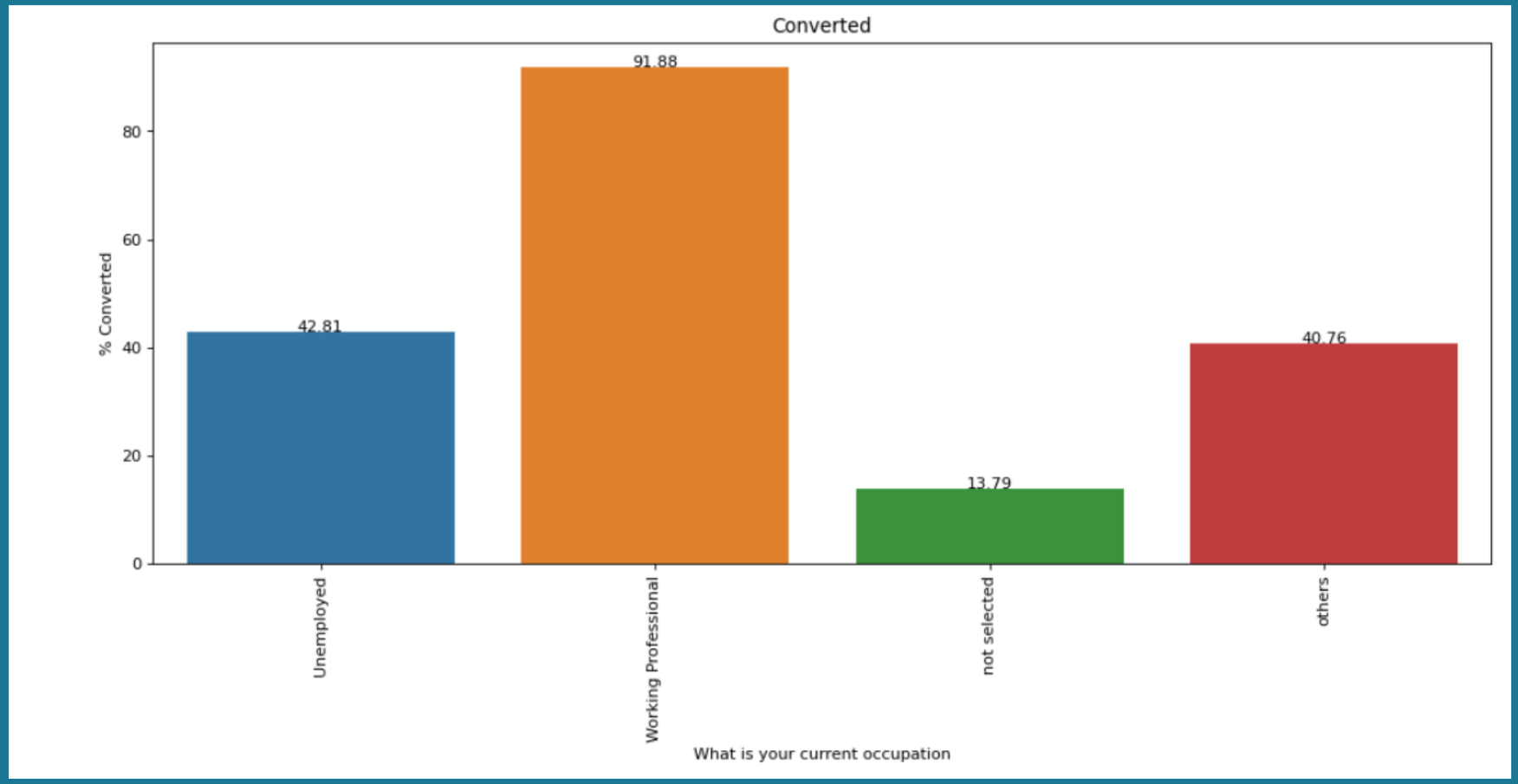
# Analysis of Last Activity




Inference :

- SMS Sent category have the highest conversion rate and number of conversions.

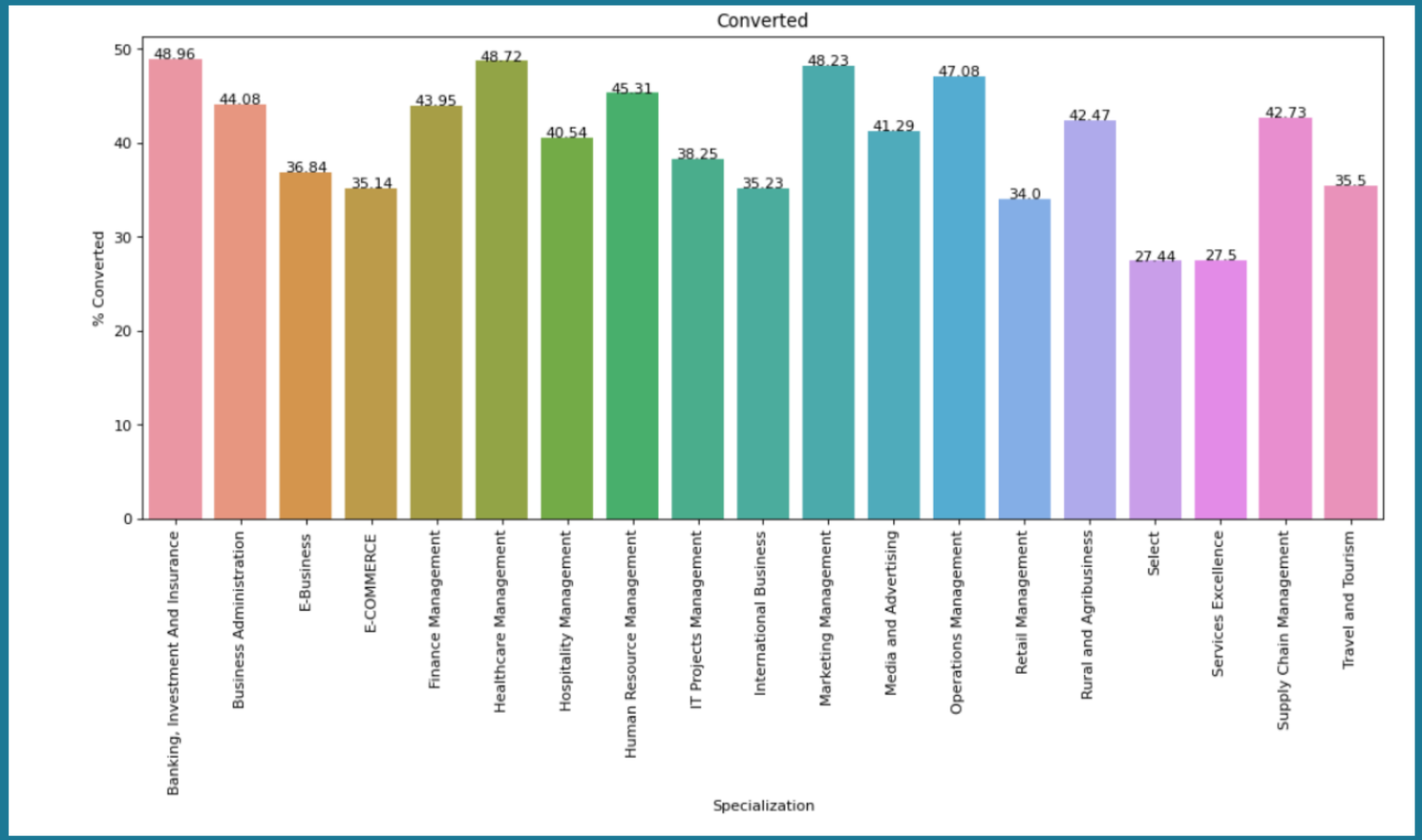
# Analysis of What is your current occupation




Inference :

- Working Professional category has heighest convertion rate.
- highest number of convertion takes place in Unemployed category.

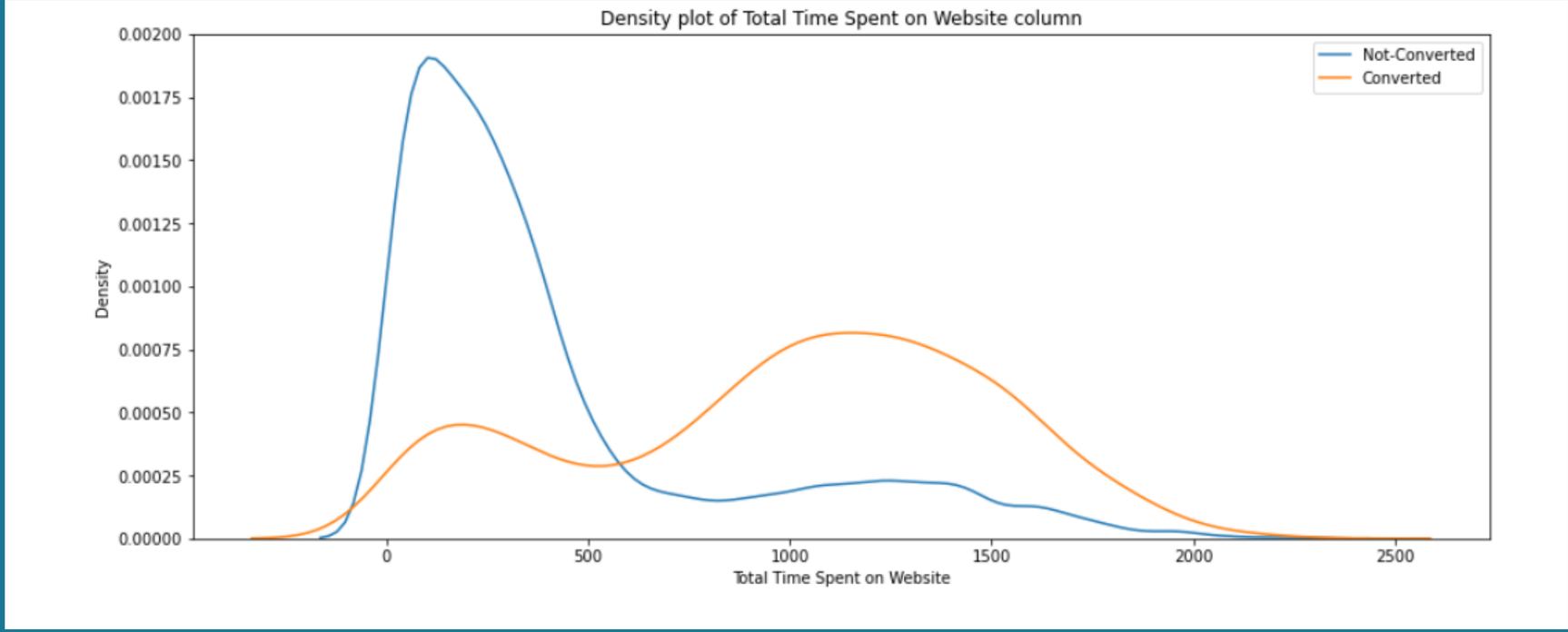
# Analysis of Specialization




Inference :

- Select category have lowest conversion rate which shows that people not selecting specialization are not intrested in the course.

# Analysis of Total Time Spent on Website




Inference :

- We can observe that as the total time spend on website increases, probablity of convertion increases.

# Final Summary

**Lets go through the inferences of each feature (arranged in descending order of there importance to the model)**

| Feature | Coefficient | Inference |
|---|---|---|
| Lead Source_Welingak Website | 6.1733 | If a lead is comming from Welingak Website it has very high chance of conversion. |
| Lead Source_Reference | 3.4076 | If a lead is comming from Reference it has very high chance of conversion. |
| What is your current occupation_Working Professional | 2.4417 | If occupation of the customer is Working Professional then it has very high chance of conversion. |
| What is your current occupation_not selected | -1.1967 | If occupation is not selected then it has very less chance of conversion. |
| Lead Source_Olark Chat | 1.1558 | If a lead is comming from Olark Chat it has very high chance of conversion. |
| Last Activity_SMS Sent | 1.1147 | If the last activity is SMS Sent then it has very high chance of conversion. |
| Total Time Spent on Website | 1.1010 | If the total time spent on the website is high then chance of conversion is higher. |
| Do Not Email | -1.0225 | If the customer is telling to not send email there is very less chance of conversion in this case. |
| Specialization_Select | -0.9376 | If Specialization is not selected then its very less chance that the lead will convert. |
| Specialization_Hospitality Management | -0.9057 | If Specialization is Hospitality Management then its very less chance that the lead will convert. |
| Lead Origin_Landing Page Submission | -0.8862 | If a lead is comming from Landing Page Submission it has very less chance of conversion |
| Last Activity_Other activity | -0.8134 | If last activity is in Other activity category then its very less chance of conversion. |
| TotalVisits_10+ | 0.5634 | If total visits are more than 10 then conversion rate is very high. |

**All these important features can also be confirmed by going back to EDA.**

# Thank You