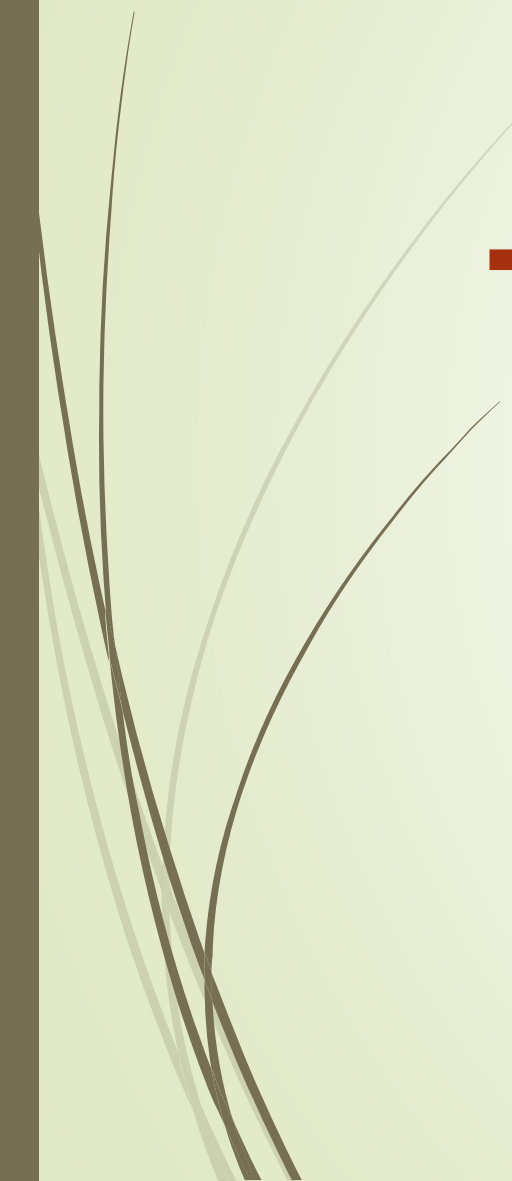




# Lead Scoring Case Study



# Goal

- The goal is to improve how we rank potential leads by taking into account different attributes like Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc., using different methods to assign scores to leads and focussing on the Hot leads for higher Lead conversion.
- 



# Problem Statement



- X Education, an online education company, caters to industry professionals by offering courses. Professionals visiting the website daily after finding the company's courses on various platforms may explore different courses and enrol in them by submitting forms. Leads are generated when visitors provide their contact details, either through forms or referrals. The sales team then contacts these leads through calls and emails, resulting in a typical conversion rate of 30%.
- Despite acquiring numerous leads, X Education struggles with a low conversion rate. To improve efficiency, the company aims to identify the most promising leads, referred to as 'Hot Leads.' By focusing efforts on these potential leads, the conversion rate is expected to increase, enhancing the effectiveness of the sales team's interactions.



# Goals of the Case Study

1. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
2. There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.



# Steps involved

- Data Sourcing
- Data Cleaning
- EDA
- Feature Scaling
- Splitting Data in Test and Training Dataset
- Model Building
- Model Evaluation
- Prediction on Test Set

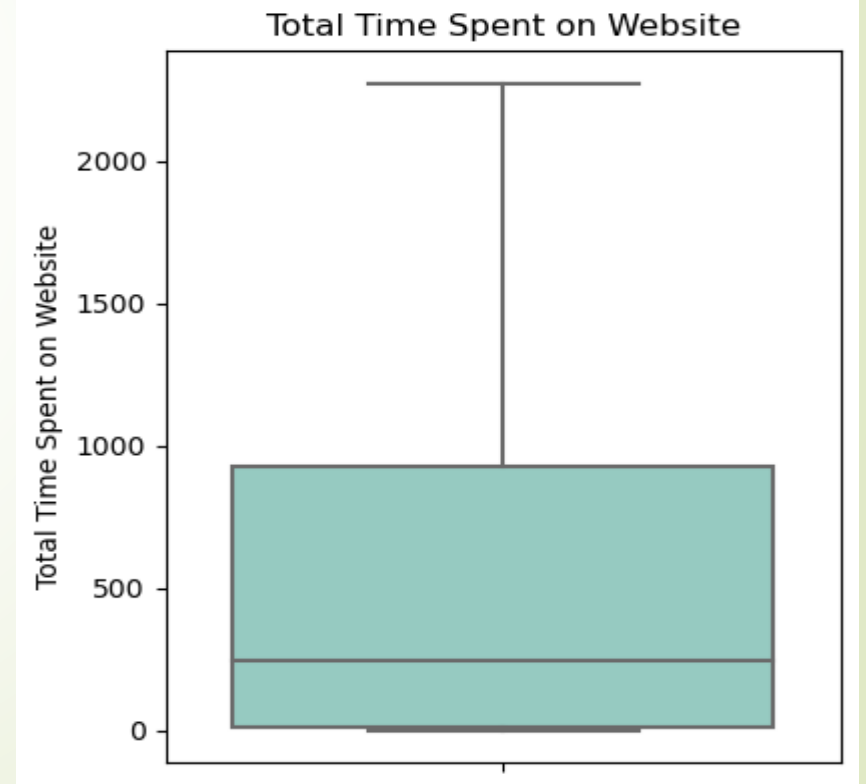
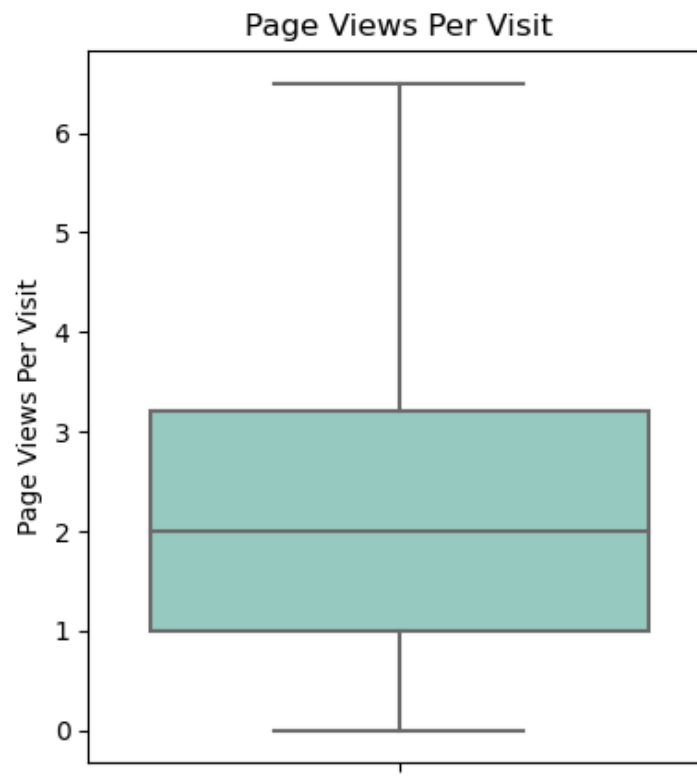
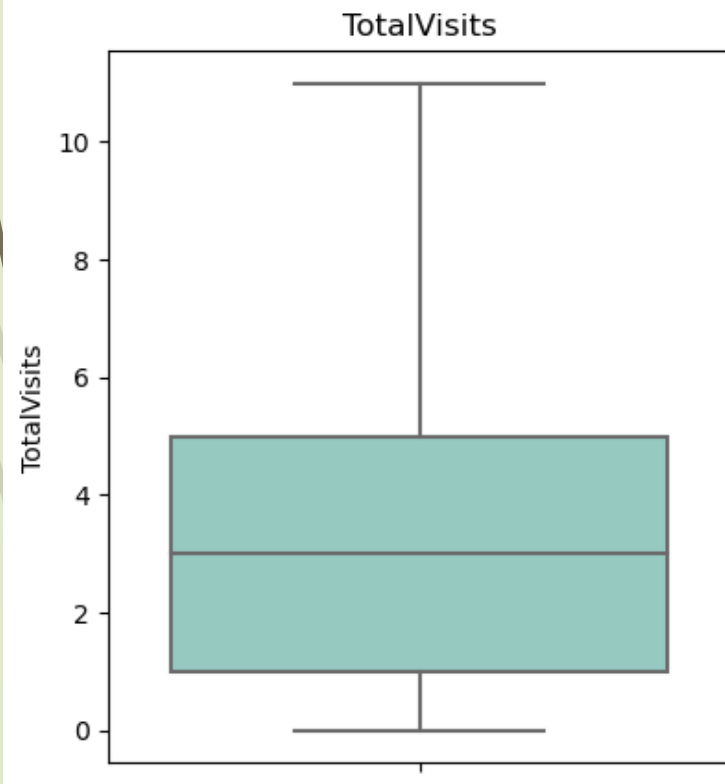


# Data Sourcing Cleaning and Preparation

- Reading the Data from the CSV file.
- Outlier Treatment
- Data Cleaning, Handling Null Values and Removing Higher Null Values Data
- Redundant Data removal from Columns
- EDA

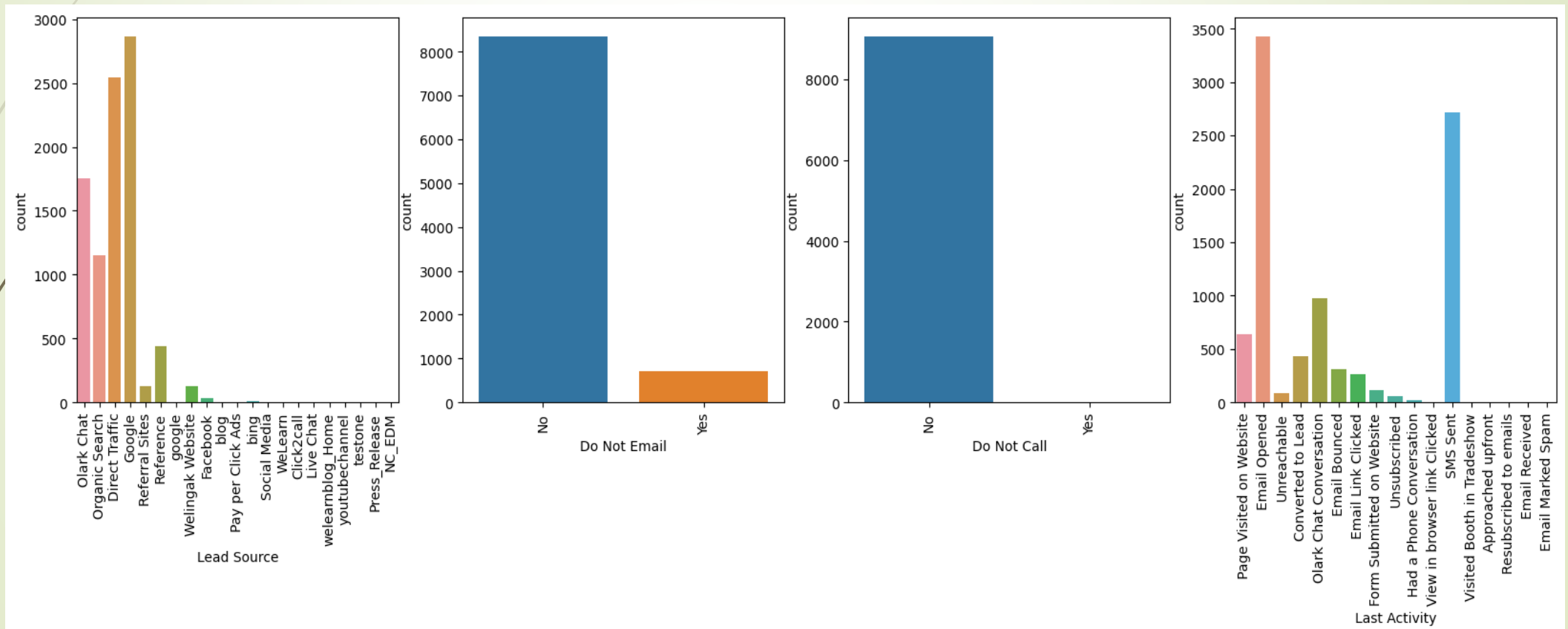
# Outliers

**Note:** to handle the outliers here 1.5 IQR rule has been applied. Any data that is greater or less than 1.5 IQR from 75th or 25th percentile respectively will be considered outlier and hence they will be capped with these values respectively



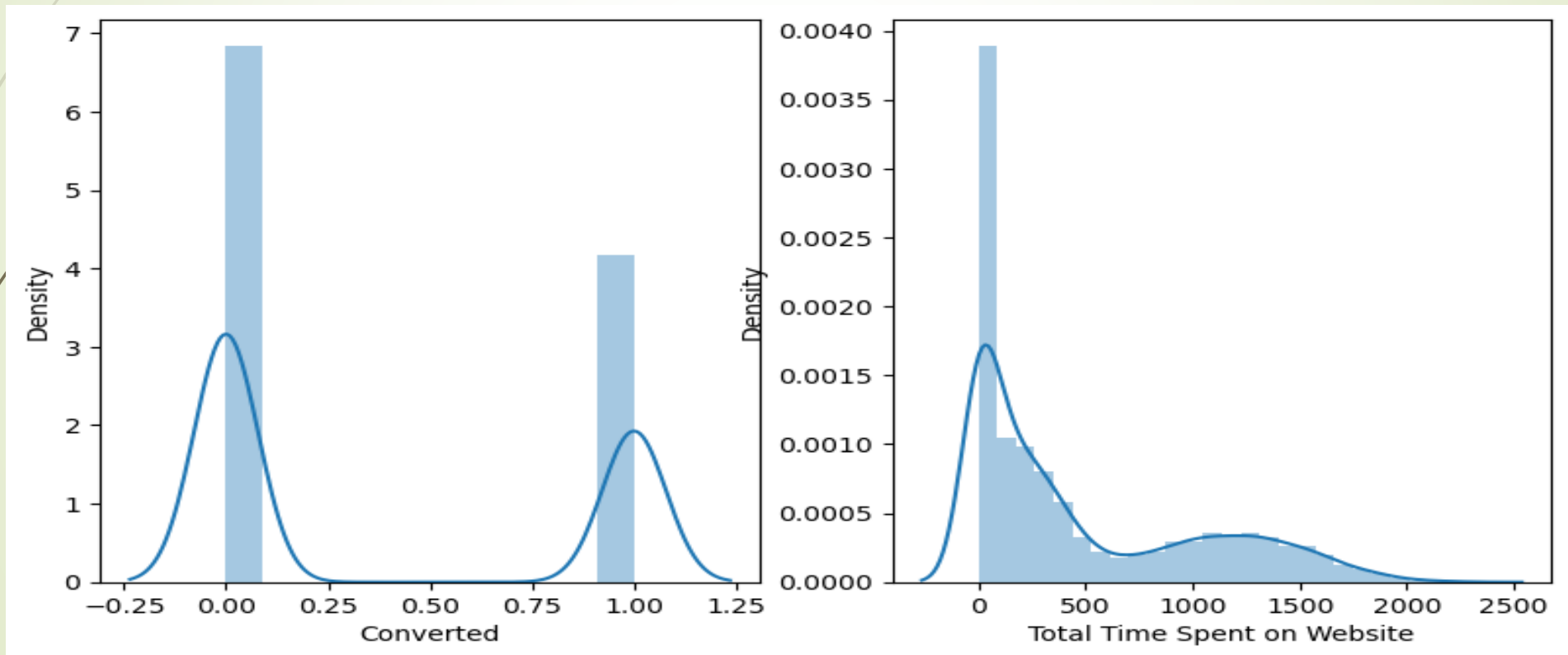


# Univariate Analysis of Categorical Variables

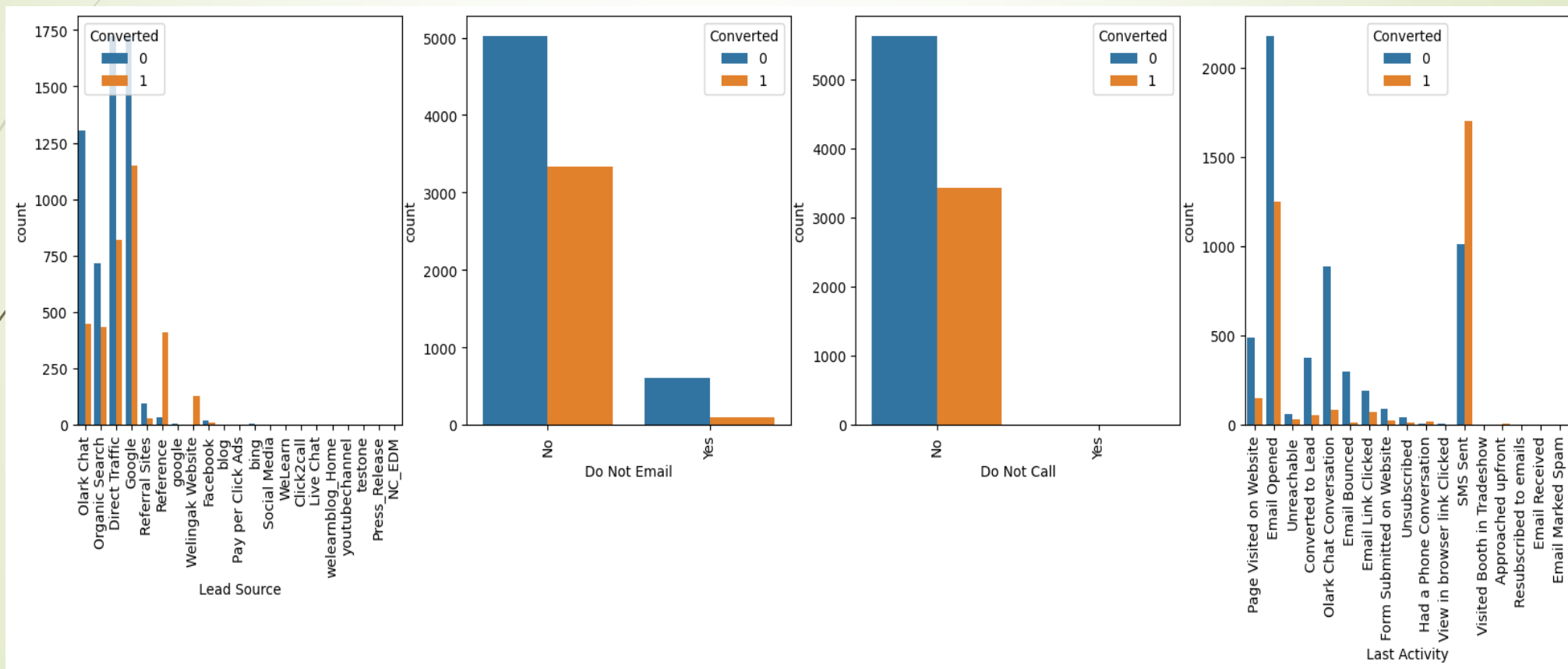




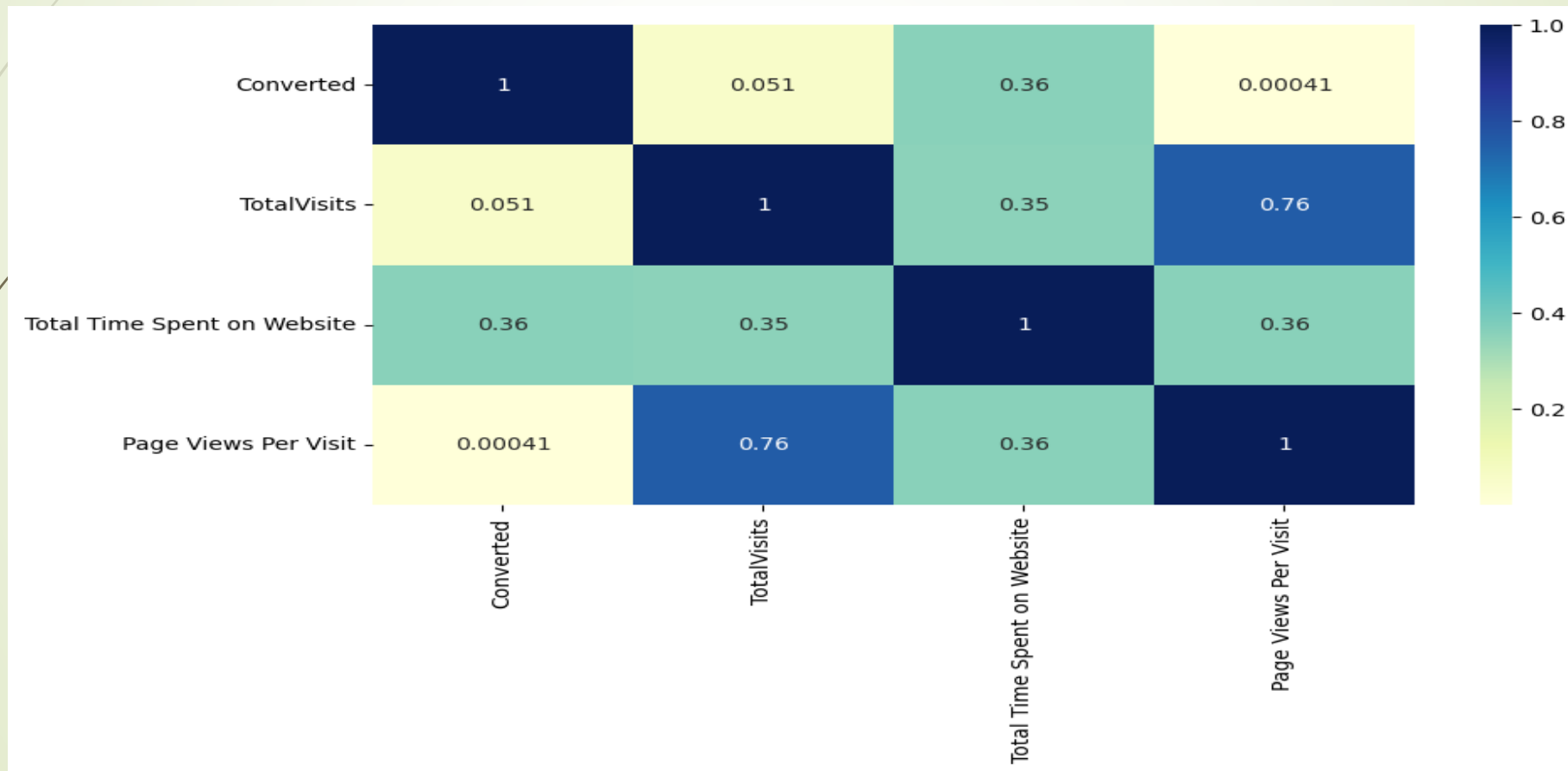
# Univariate Analysis of Numerical Variables



# Bivariate Analysis




# Multivariate Analysis Using Heatmaps





# Model Building

1. Feature Selection using RFE
  2. Determined Optimal Model using Logistic Regression
  3. Accuracy, Sensitivity and Specificity Calculation
- 

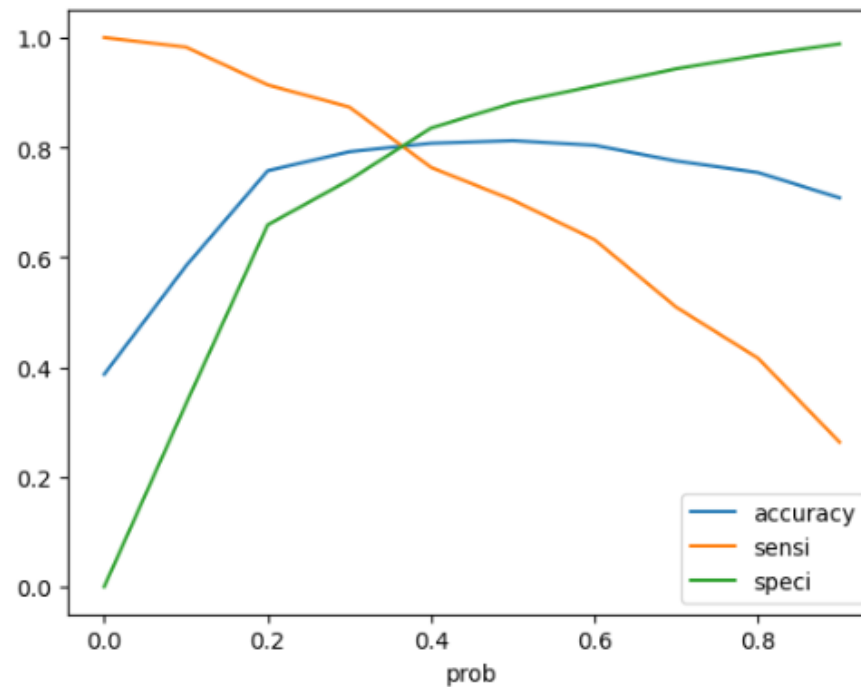


# Impact Variables

1. Total Visits
2. Total Time Spent on website
3. Page views per visit
4. Lead Origin\_Landing Page Submission
5. Lead Origin\_Lead Add Form
6. Lead Origin\_Lead Import
7. Lead Source\_Direct Traffic
8. Lead Source\_Facebook
9. Lead Source\_Google
10. Last Notable Activity\_Modified etc.

# Model Evaluation-Sensitivity and Specificity on Training Data Set

```
In [590]: cutoff_df.plot.line(x='prob', y=['accuracy', 'sensi', 'speci'])  
plt.show()
```



Selecting Cutoff as 0.38  
from graph based on:  
1. Accuracy = 80.55%  
2. Specificity = 82.46% and  
3. Sensitivity = 77.52%



# Result

- The model achieves a reasonably good accuracy, sensitivity, and specificity, suggesting that it is capable of identifying hot leads to a certain extent.
- - The accuracy score is around 80-81%, indicating that the model's predictions are accurate for a significant portion of the dataset.
- - The sensitivity score is also around 80%, which means that the model is able to correctly identify a substantial proportion of actual converting leads.
- - The specificity score is around 81%, implying that the model can effectively distinguish between converting and non-converting leads.


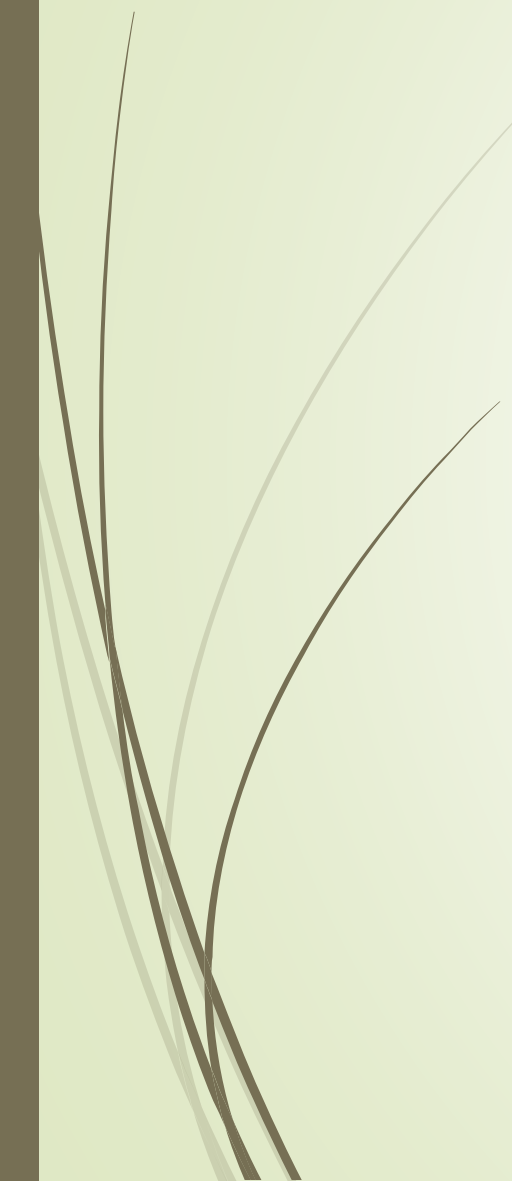




# Recommendations



- Leads who spend more time on the website are more likely to be hot leads. This indicates their higher engagement and interest in the offerings.
- Leads who come from the 'Welingak Website' as the lead source are more likely to convert. This could be due to the effectiveness of this source in attracting genuinely interested leads.
- Leads with the last activity being 'SMS Sent' are more likely to convert. This suggests that following up with leads through SMS communication is effective in driving conversions.

- 
- 
- Leads who are 'Working Professionals' have a higher likelihood of converting. This indicates that individuals with a specific professional background are more likely to become customers.
  - Leads associated with the 'Modified' last notable activity have a higher conversion probability. This might indicate that personalized follow-ups contribute to higher conversions.