Lead Scoring Case Study

Group:

Siva Gopavarapu

Manan Sharma

Saheli Paul

Background

- X Education , An education company named sells online courses to industry professionals
- The company markets its courses on several websites 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
- 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%

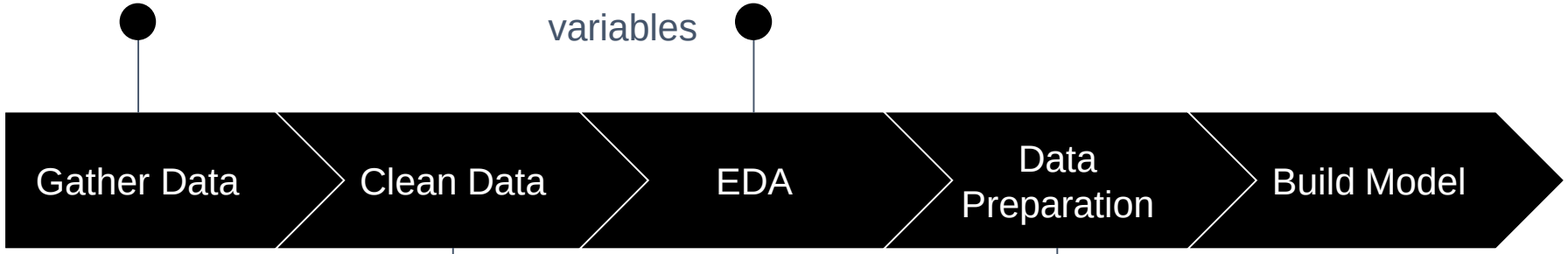
Problem Statement

- 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
- Propose the changes which will help to increase the conversion rate
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be 80%.

Case Study Implementation

Get the data provided by the company

Perform the data analysis and build graphs to understand the relation between variables

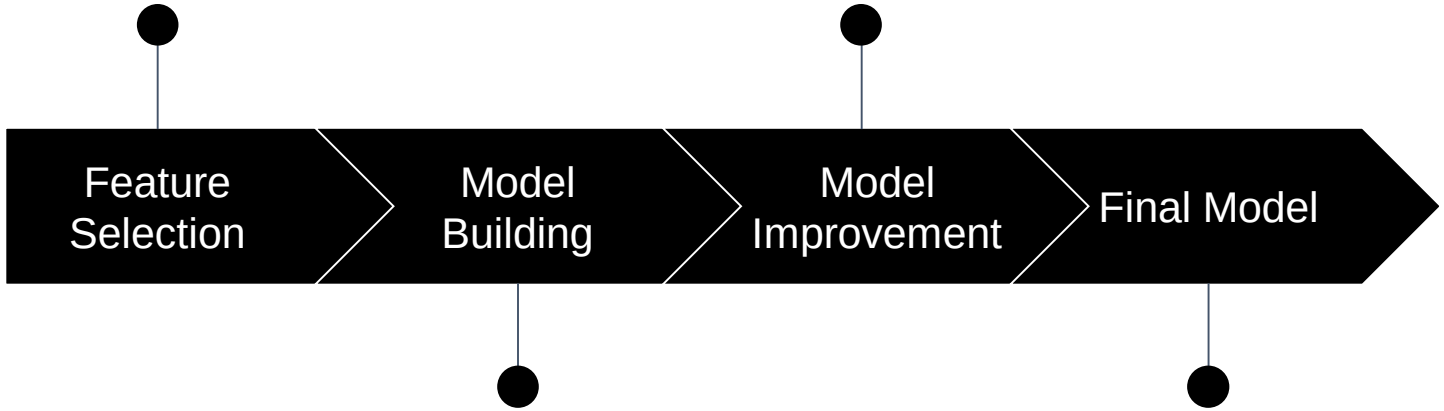


Clean the data such as null value treatment, outliers

Scale the numerical columns and use rfe for variable selection

Selection of top 15 features using RFE

Check the p value and vif to change the columns for model

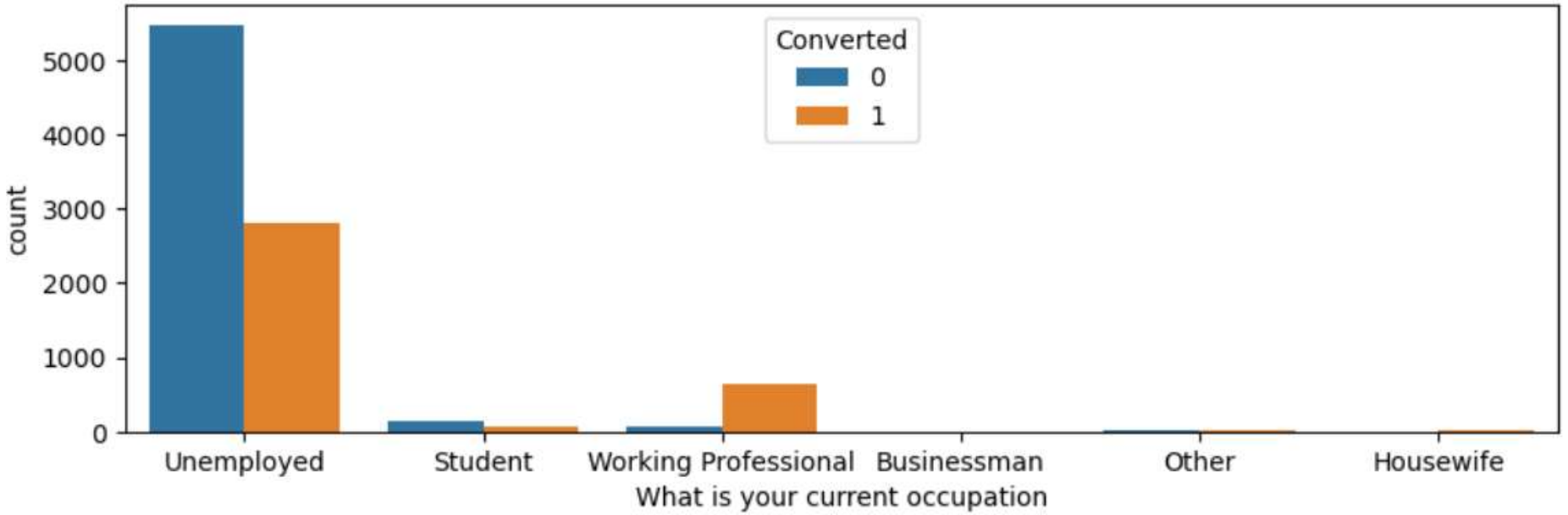


Model building using RFE for selected columns

Analyse the final model based on accuracy, sensitivity, and specificity

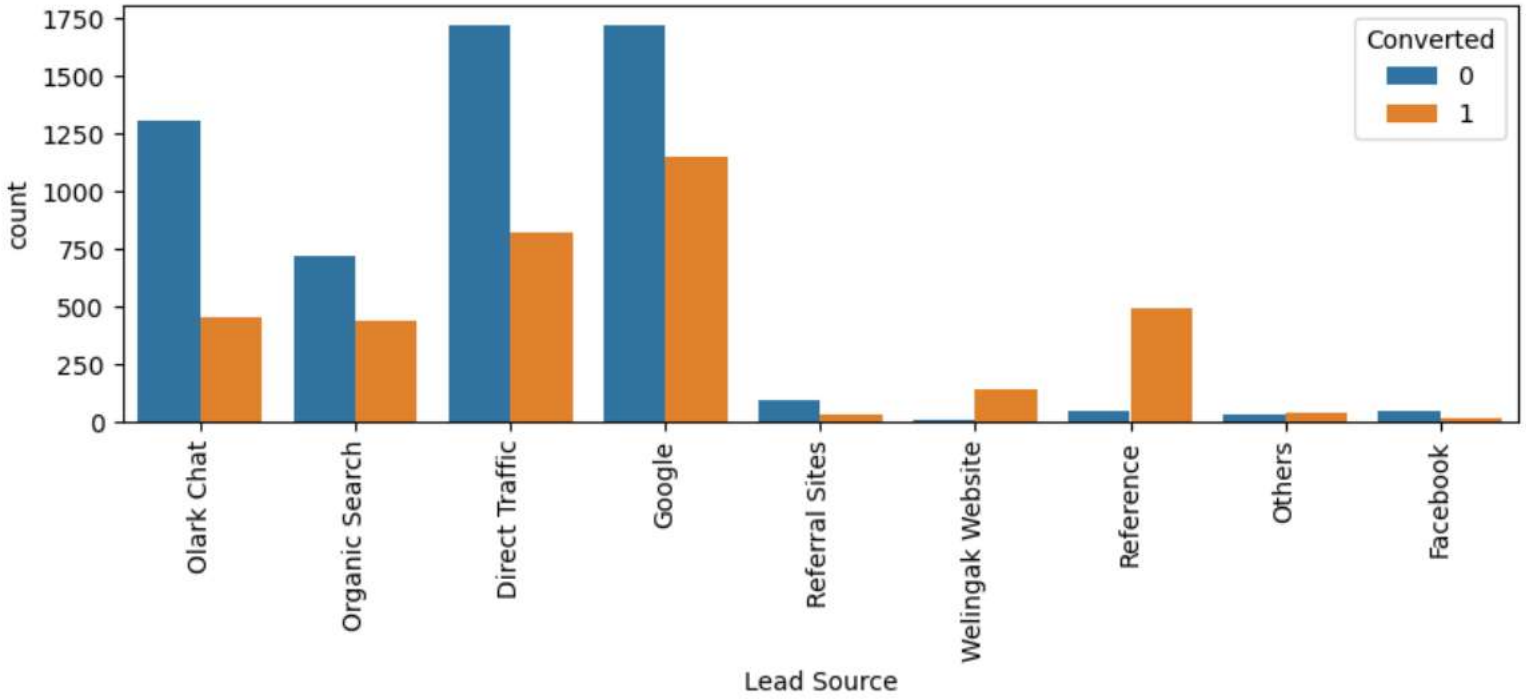
Graphs from EDA

Occupation Vs Converted



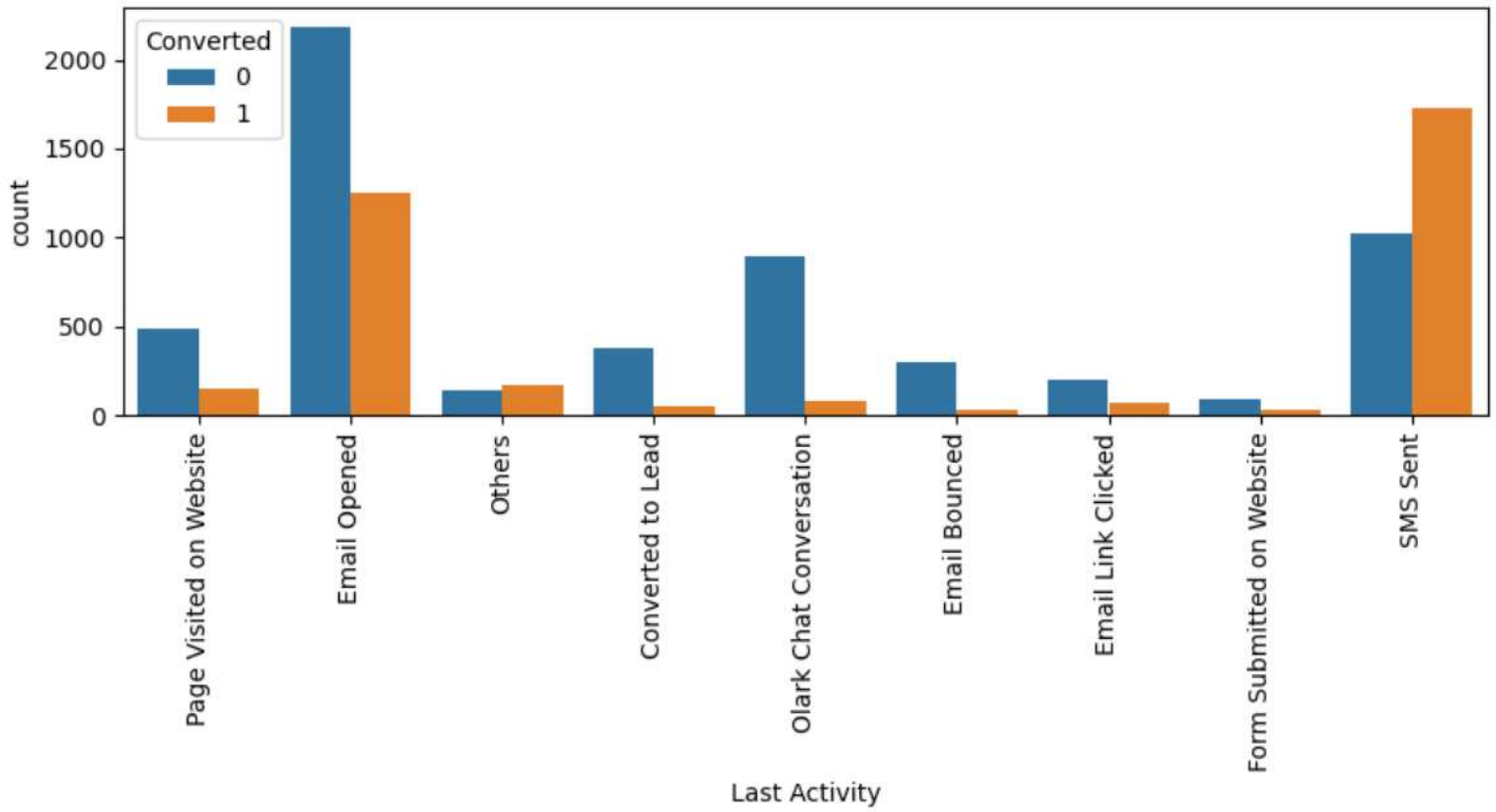
It’s clear that working professionals are having very high chance of conversion

Lead Source Vs Converted



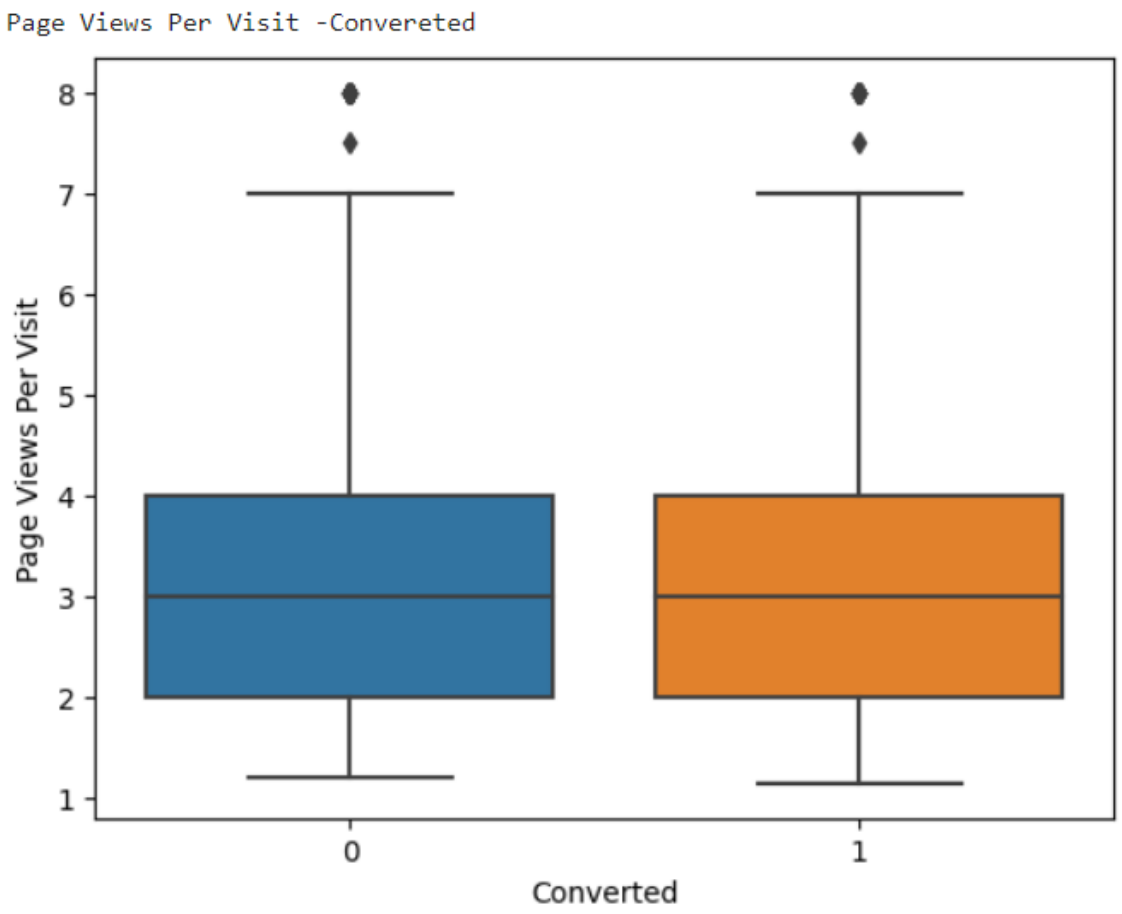
Proportion of leads converted is highest in reference followed by welingak_website
Maximum number of leads are from direct traffic and google
Organic search also has good conversion rate
Concentrate on increasing the reference and organic search lead base for more conversion

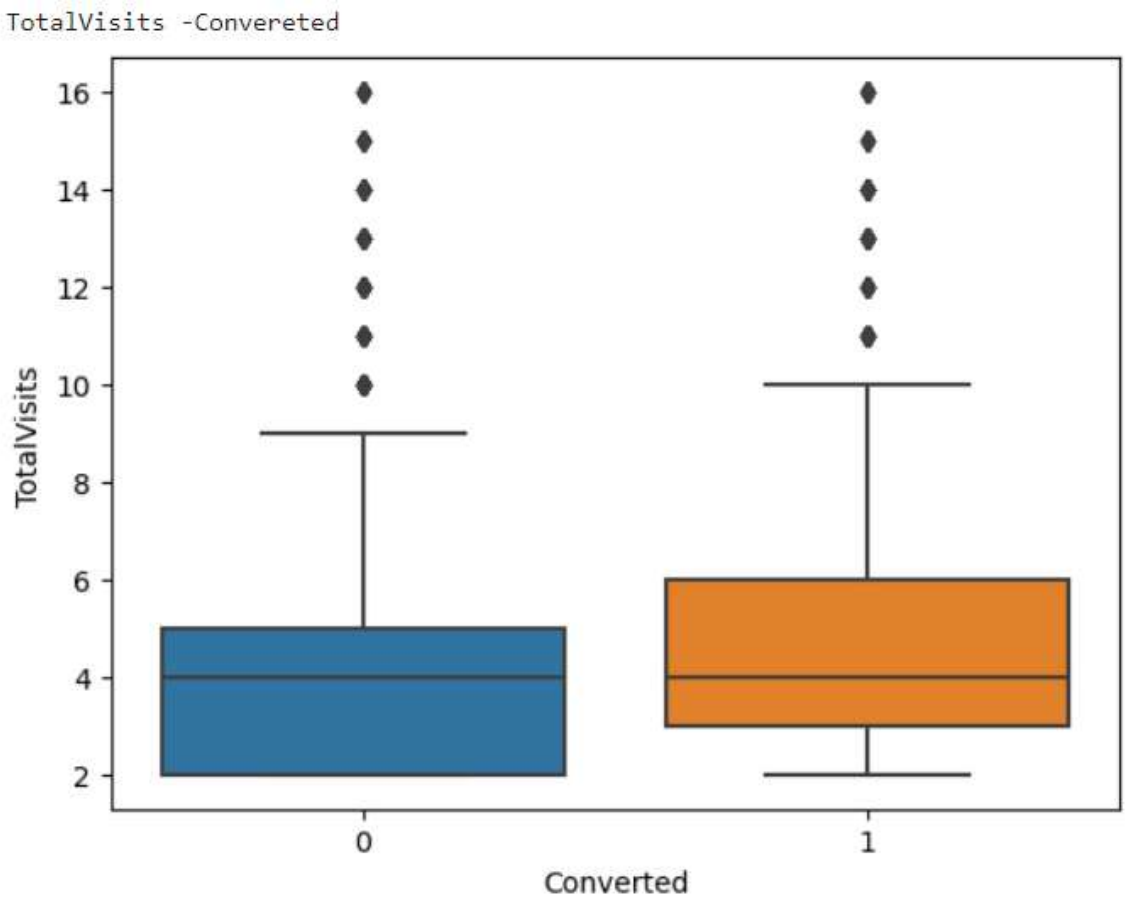
Last Activity Vs Converted



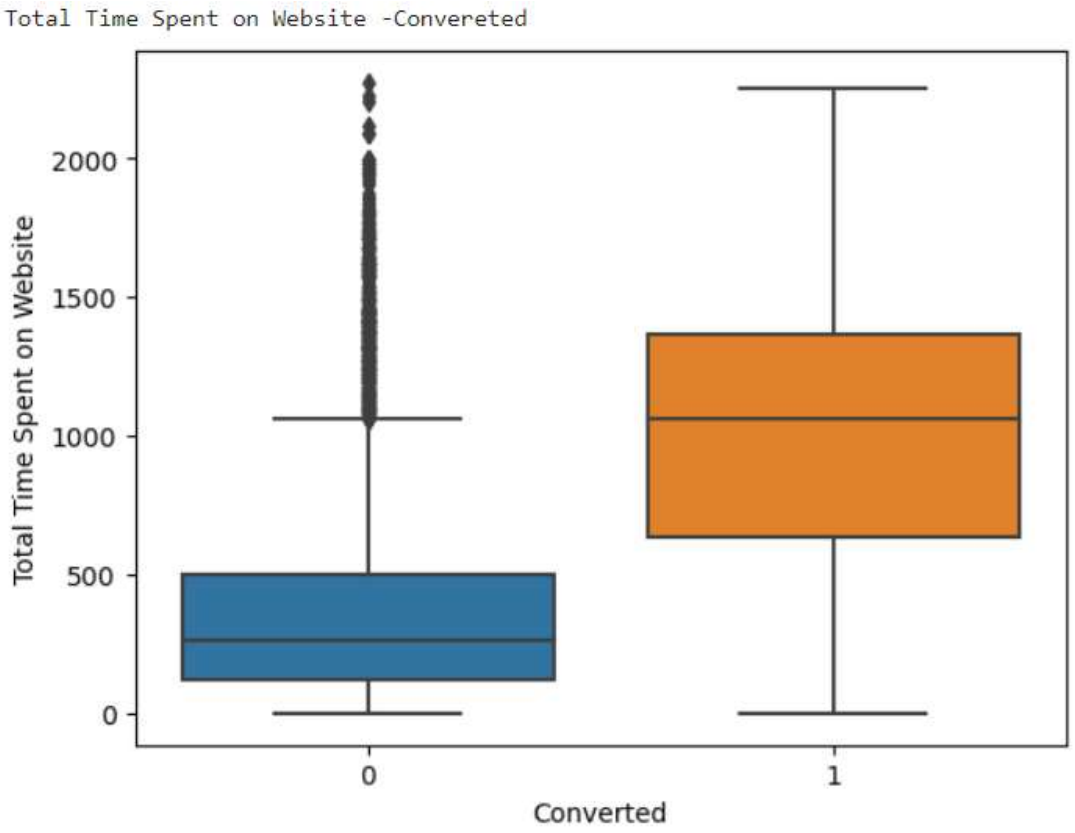
SMS is the last interaction channel where the lead conversion is the highest followed by email opened

Pages View Per Visit and Total Visits Vs Converted



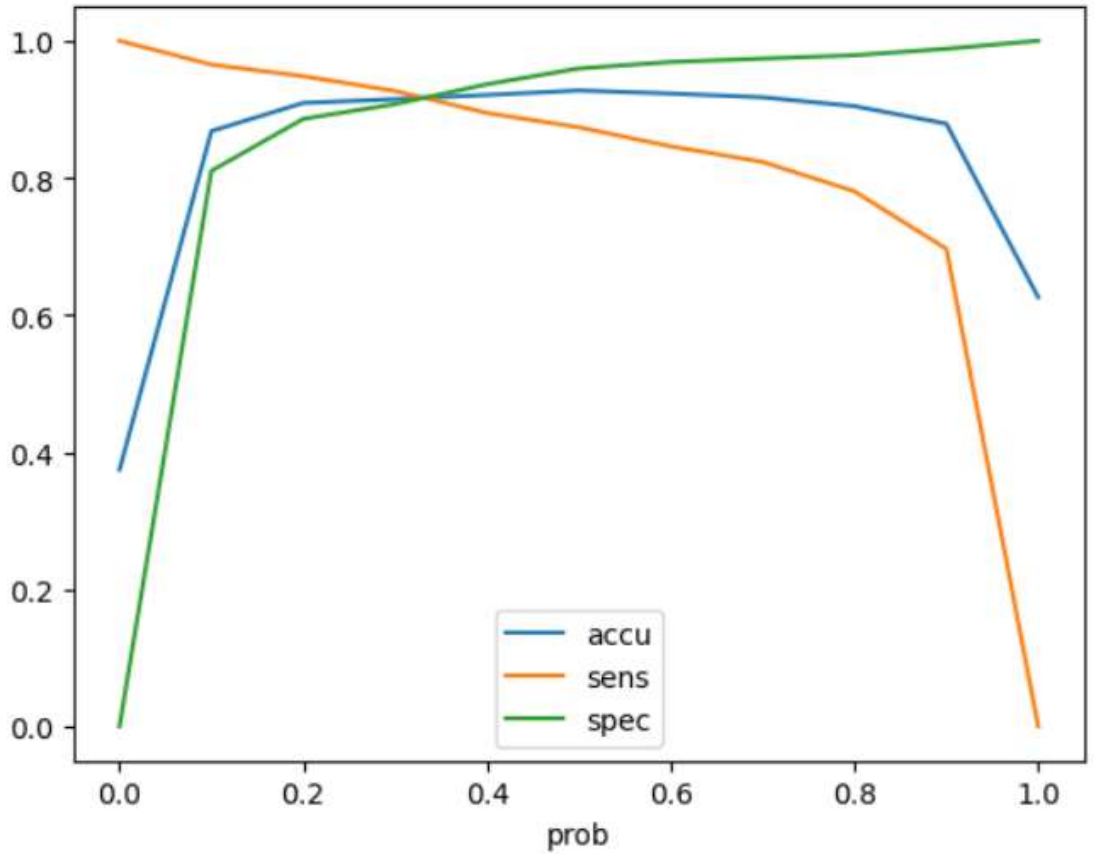


Total Time Spent on Website



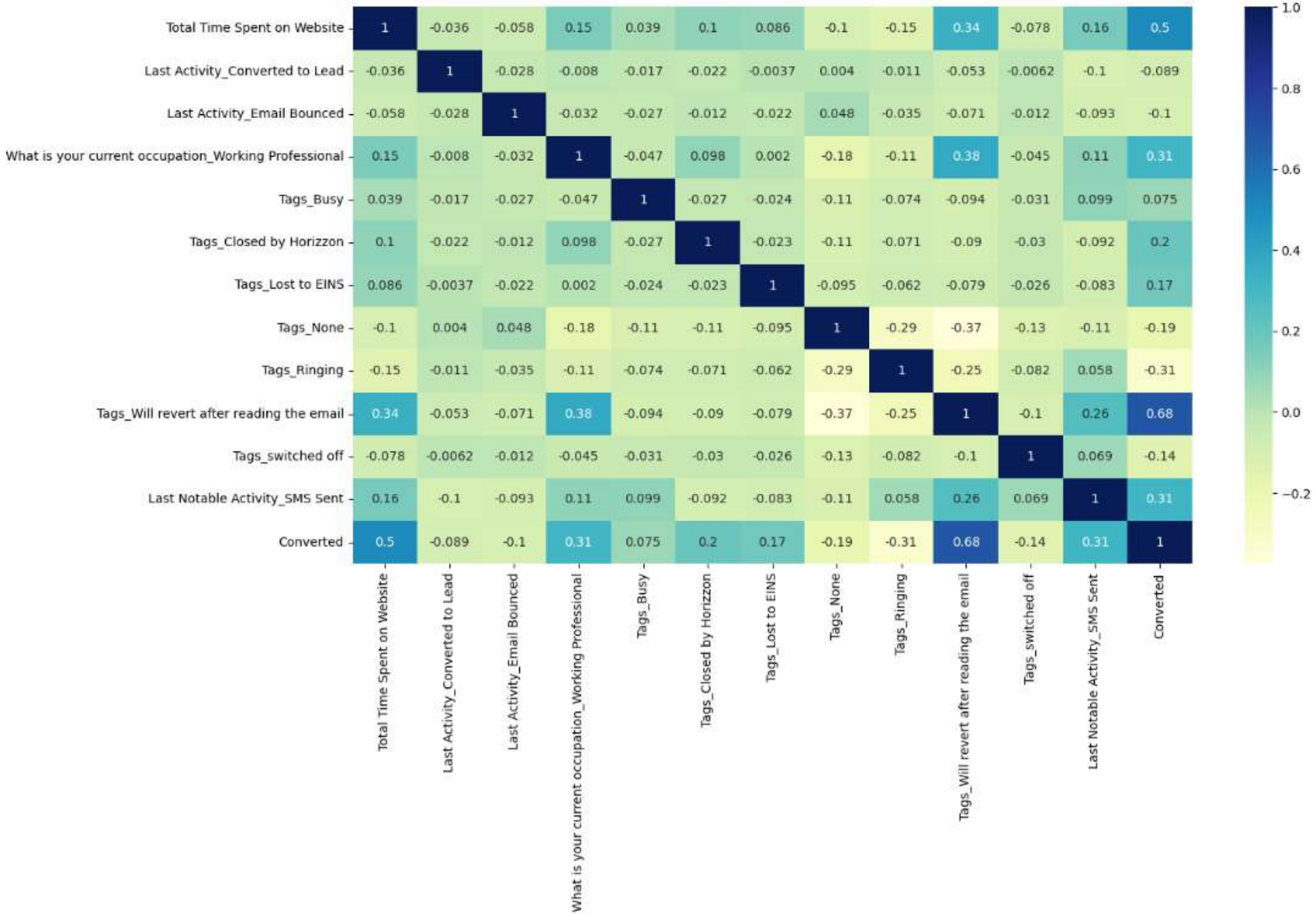
Leads who spend more time on website have higher change of conversion

Select Optimum Cutoff Value



0.3 is the optimum cutoff value for the LogisticRegression model

Heatmap to see the correlation



Conclusion from the model

Top 3 predictor variables:

1. Tags_Will revert after reading the email
2. Total Time Spent on Website
3. What is your current occupation_Working Professional / Last Notable Activity_SMS Sent

Train and Test dataset metrics

Training:

- Accuracy: 91.48%
- Sensitivity: 92.68%
- Specificity: 90.76%

Testing:

- Accuracy: 91.57%
- Sensitivity: 92.11%
- Specificity: 91.24%