Lead Score Case Study

ABHIJEETA BALIYARSINGH
SANJU RAJAN P
DEEPAK CHAKALA

Problem Statement:

➤ An X Education company has a very poor lead conversion rate.

➤ The company wishes to identify the most potential leads, also known as 'Hot Leads', so that it can improve the conversion rate.

➤ The company requires us 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.

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

Business Goals & Objective :

There are quite a few goals for this case study.

➢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 into a paying customer whereas a lower score would mean that the lead is cold and will mostly not get converted.

➢The target lead conversion rate need to be around 80%.

Solution Methodology :

➢ **Data cleaning and data manipulation**.

1. Drop the duplicate records.

2. Drop columns, if it contains large amount of missing values and not useful for the analysis.

3. Handle the missing values. Imputation of the values, if necessary.

4. Check and handle outliers in data.

➤EDA

1. Univariate data analysis: value count, data distribution
2. Segmented Univariate data analysis.
3. Bivariate data analysis: correlation coefficients and pattern between the variables etc.
4. Multivariate Data Analysis

➤Dummy Variables Creations Where needed

➤ Scaling , Feature Selection and encoding of the data.

➤Classification technique: logistic regression used for the model making and prediction.
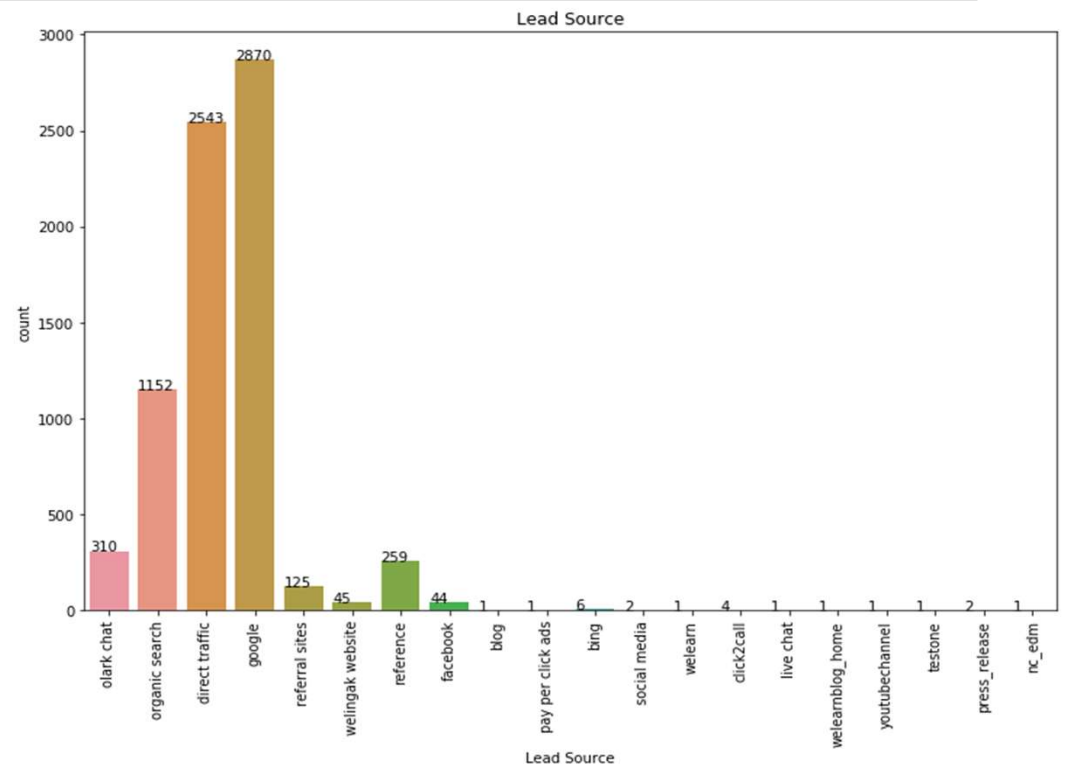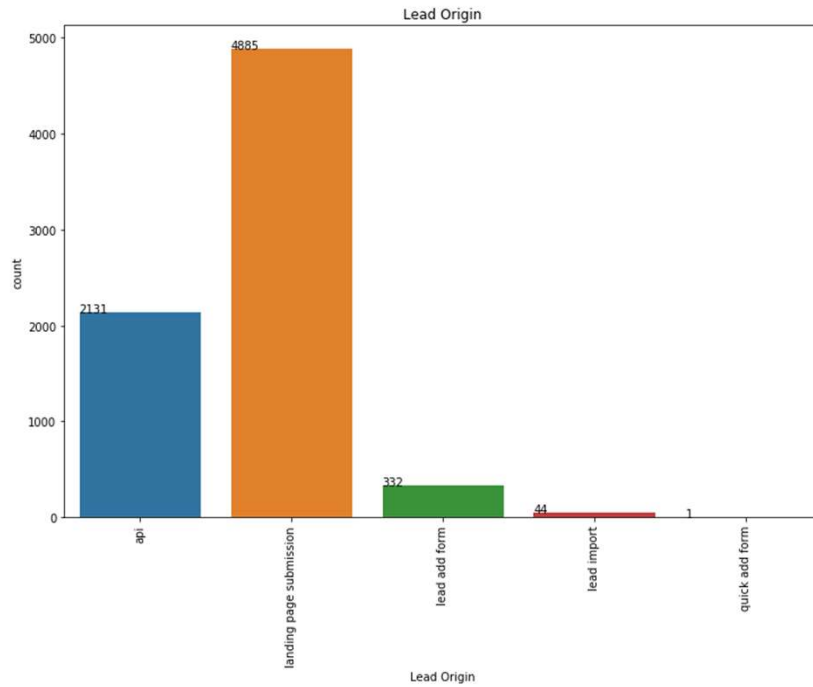
➤Validation of the model.

➤Model presentation.

➤Conclusions and recommendations.

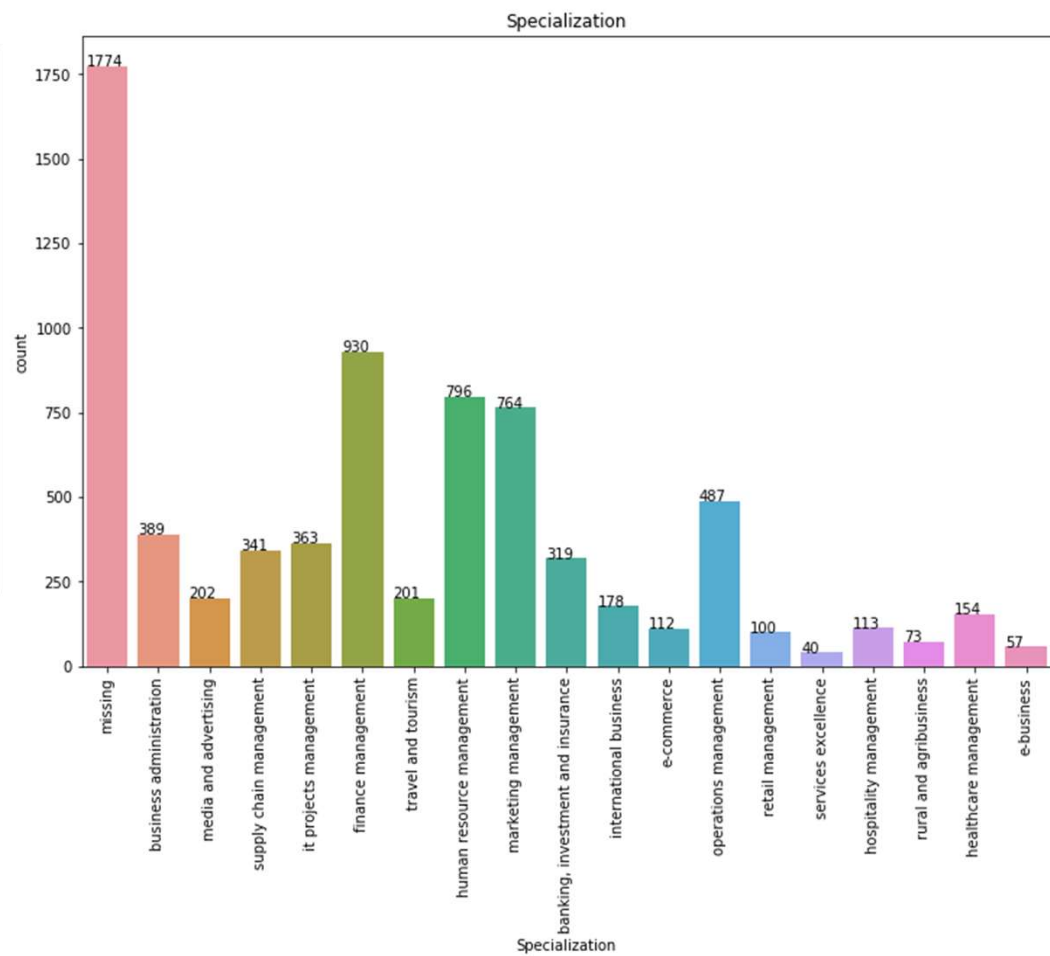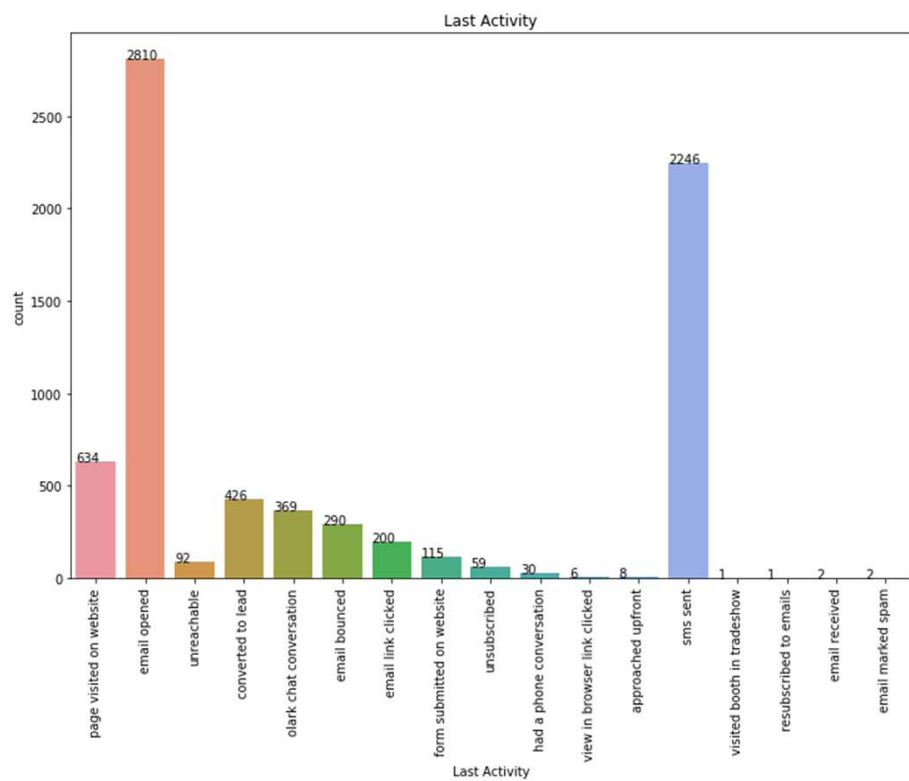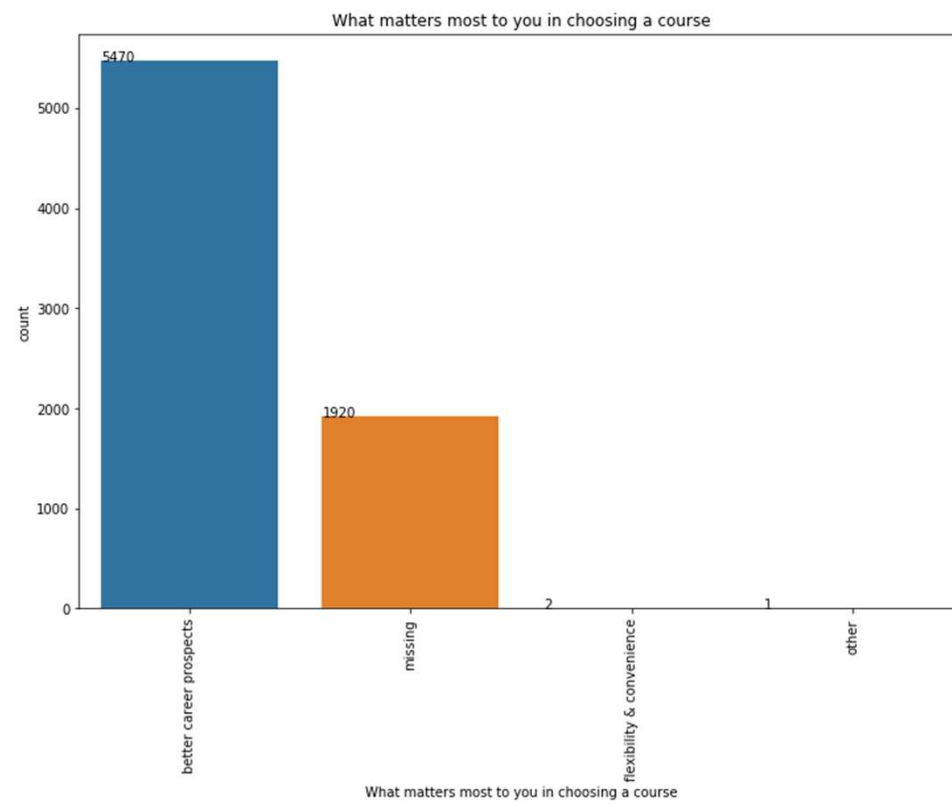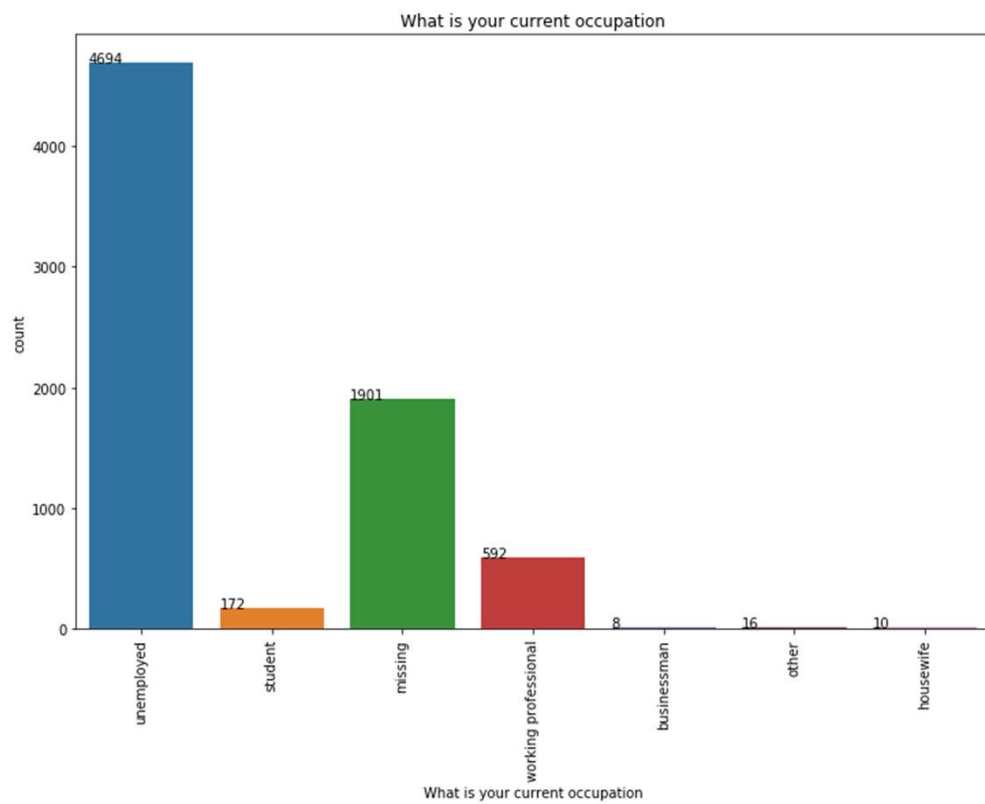# Data Manipulation :

➤ Total Number of Rows = 9240, Total Number of Columns = 37

➤ Single value features like "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" etc. have been dropped.

➤ Removing the "Prospect ID" and "Lead Number" which is not necessary for the analysis

➤ Dropping the columns having more than 30% as missing value such as 'How did you hear about X Education' and 'Lead Profile'.
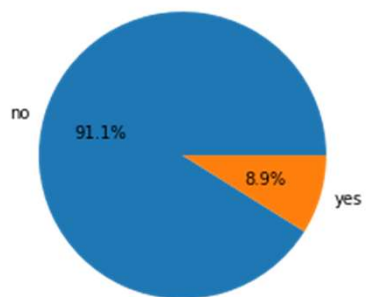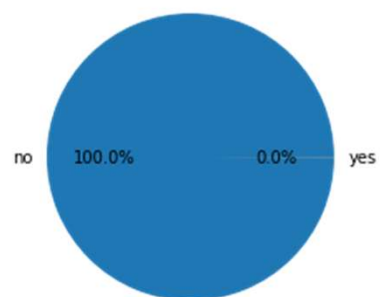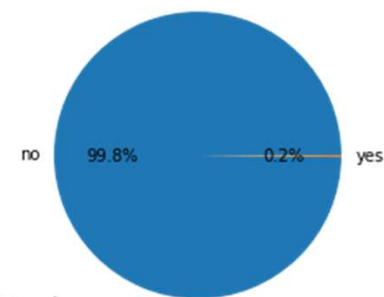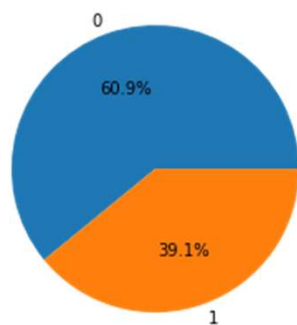
EDA :

Last Activity

Last Activity	count
page visited on website	634
email opened	2810
unreachable	92
converted to lead	426
olark chat conversation	369
email bounced	290
email link clicked	200
form submitted on website	115
unsubscribed	59
had a phone conversation	30
view in browser link clicked	6
approached upfront	8
sms sent	2246
visited booth in tradeshow	1
resubscribed to emails	1
email received	2
email marked spam	2

Specialization

Specialization	count
missing	1774
business administration	389
media and advertising	202
supply chain management	341
it projects management	363
finance management	930
travel and tourism	201
human resource management	796
marketing management	764
banking, investment and insurance	319
international business	178
e-commerce	112
operations management	487
retail management	100
services excellence	40
hospitality management	113
rural and agribusiness	73
healthcare management	154
e-business	57

What is your current occupation

Bar chart values (left to right): unemployed 4694, student 172, missing 1901, working professional 592, businessman 8, other 16, housewife 10

What matters most to you in choosing a course

Bar chart values (left to right): better career prospects 5470, missing 1920, flexibility & convenience 2, other 1
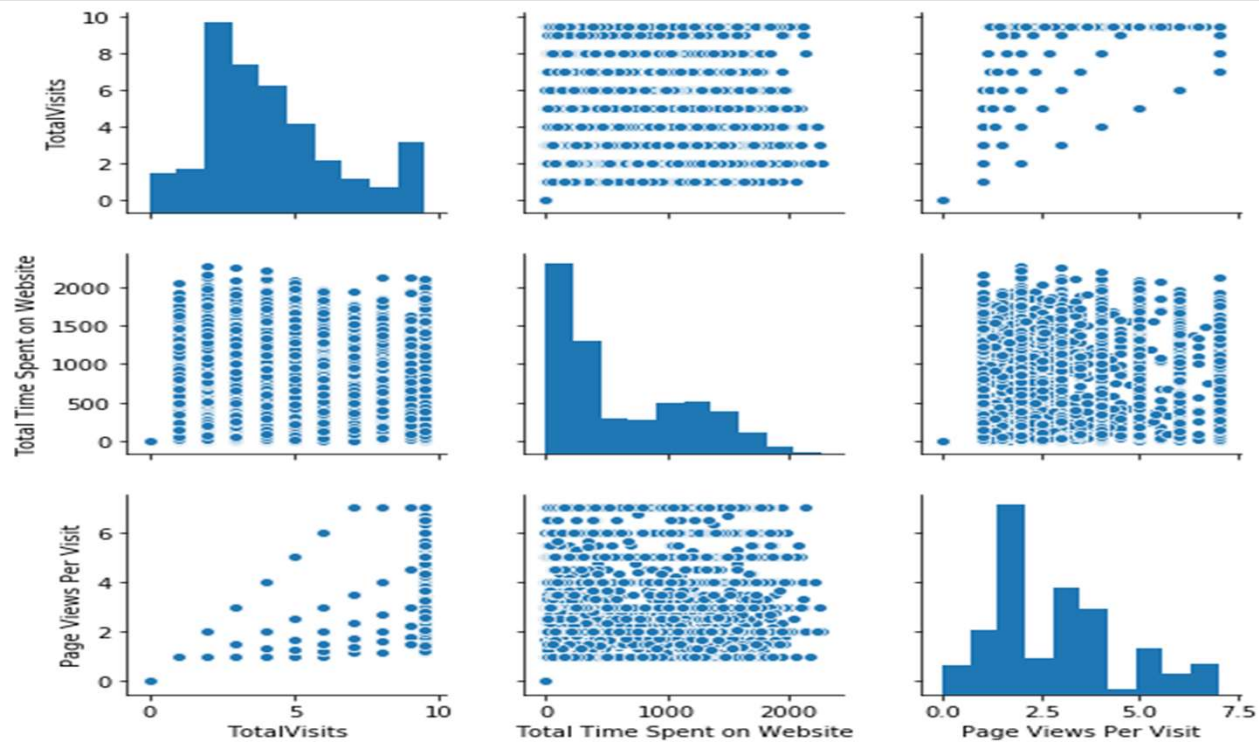
Categorical Variable Relation :

Specialization
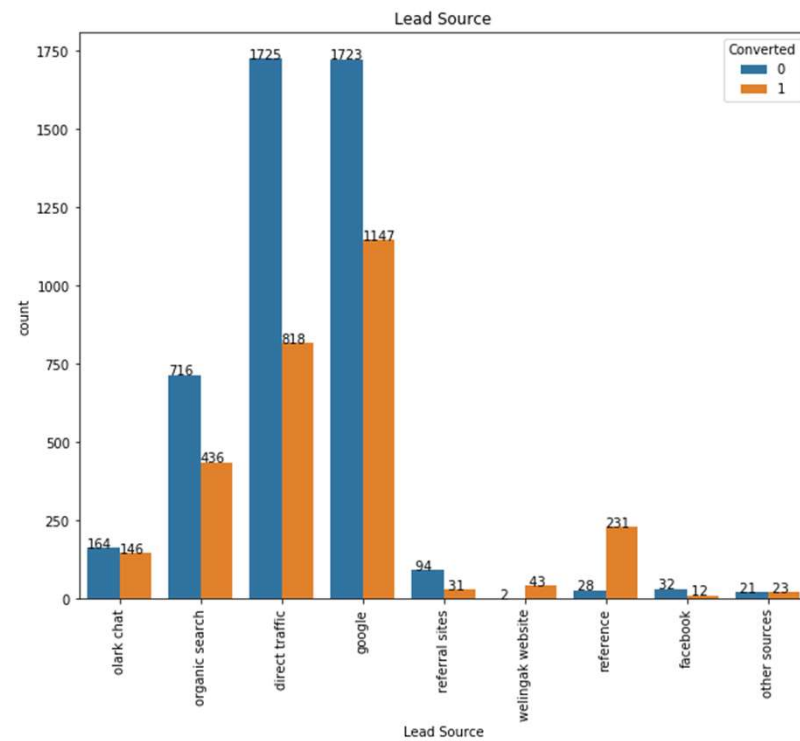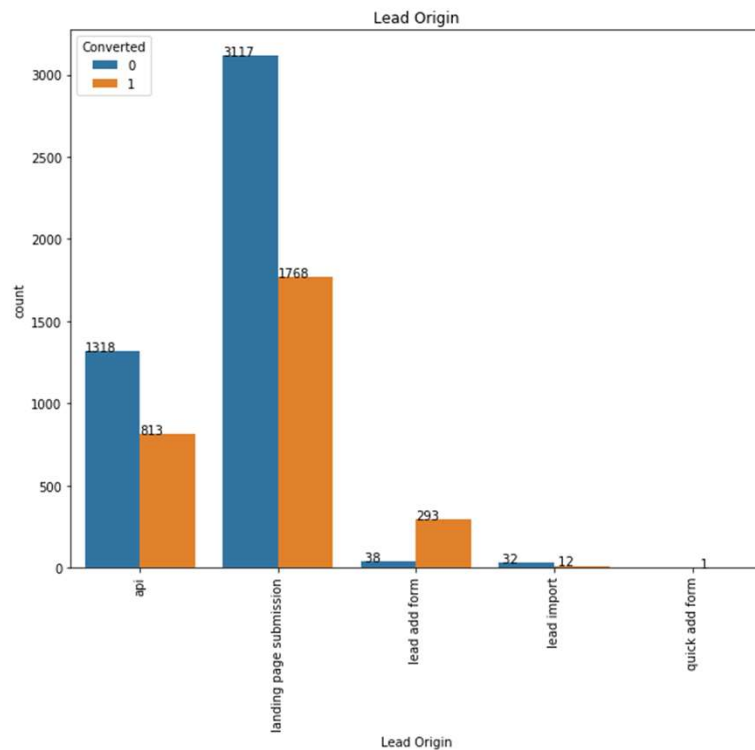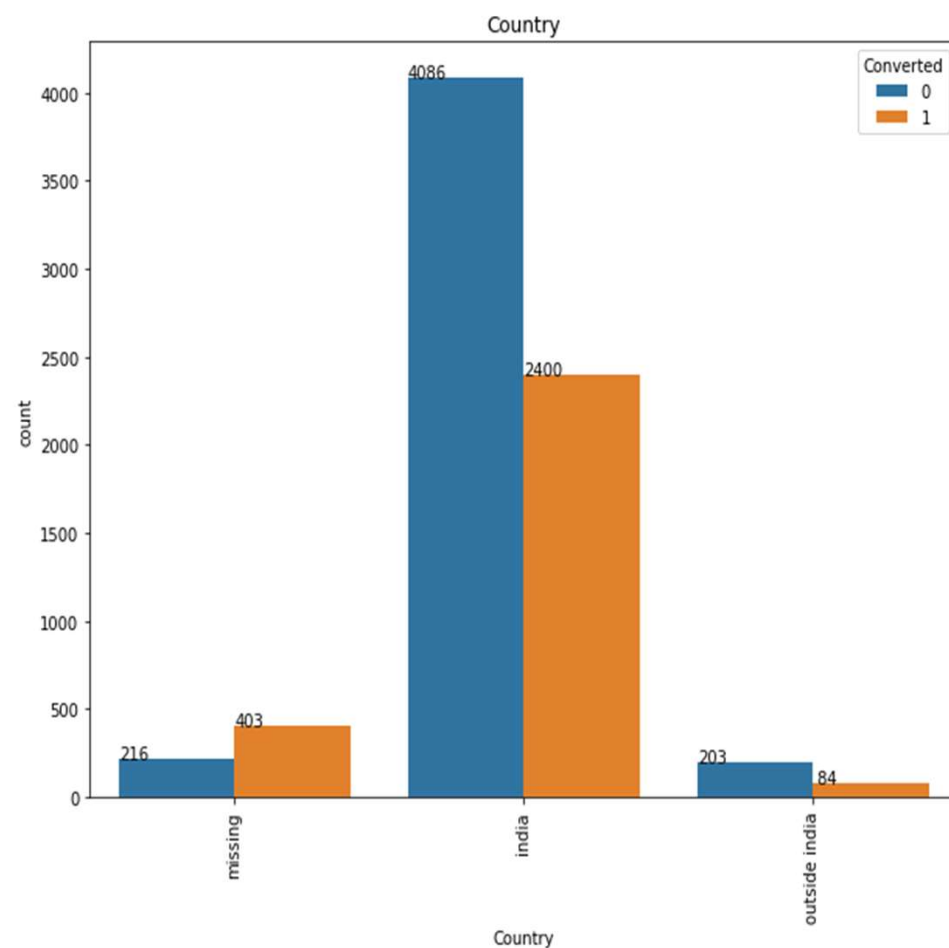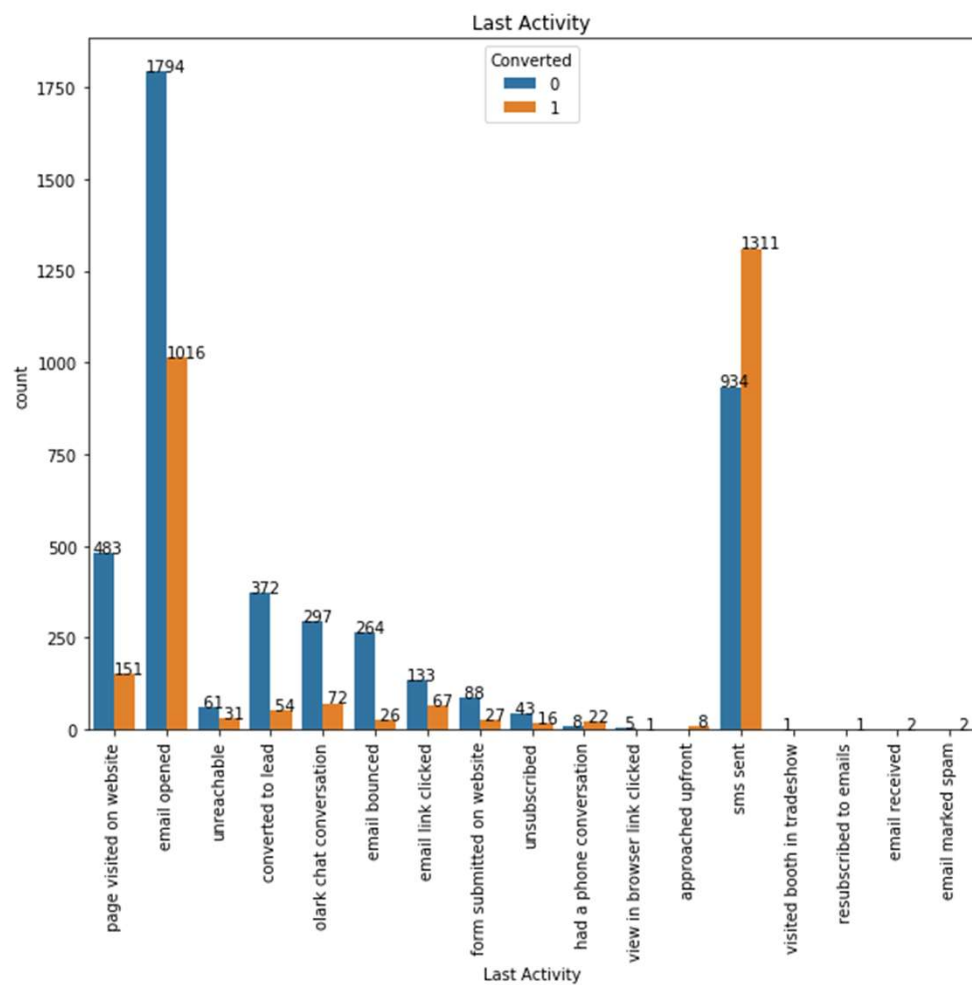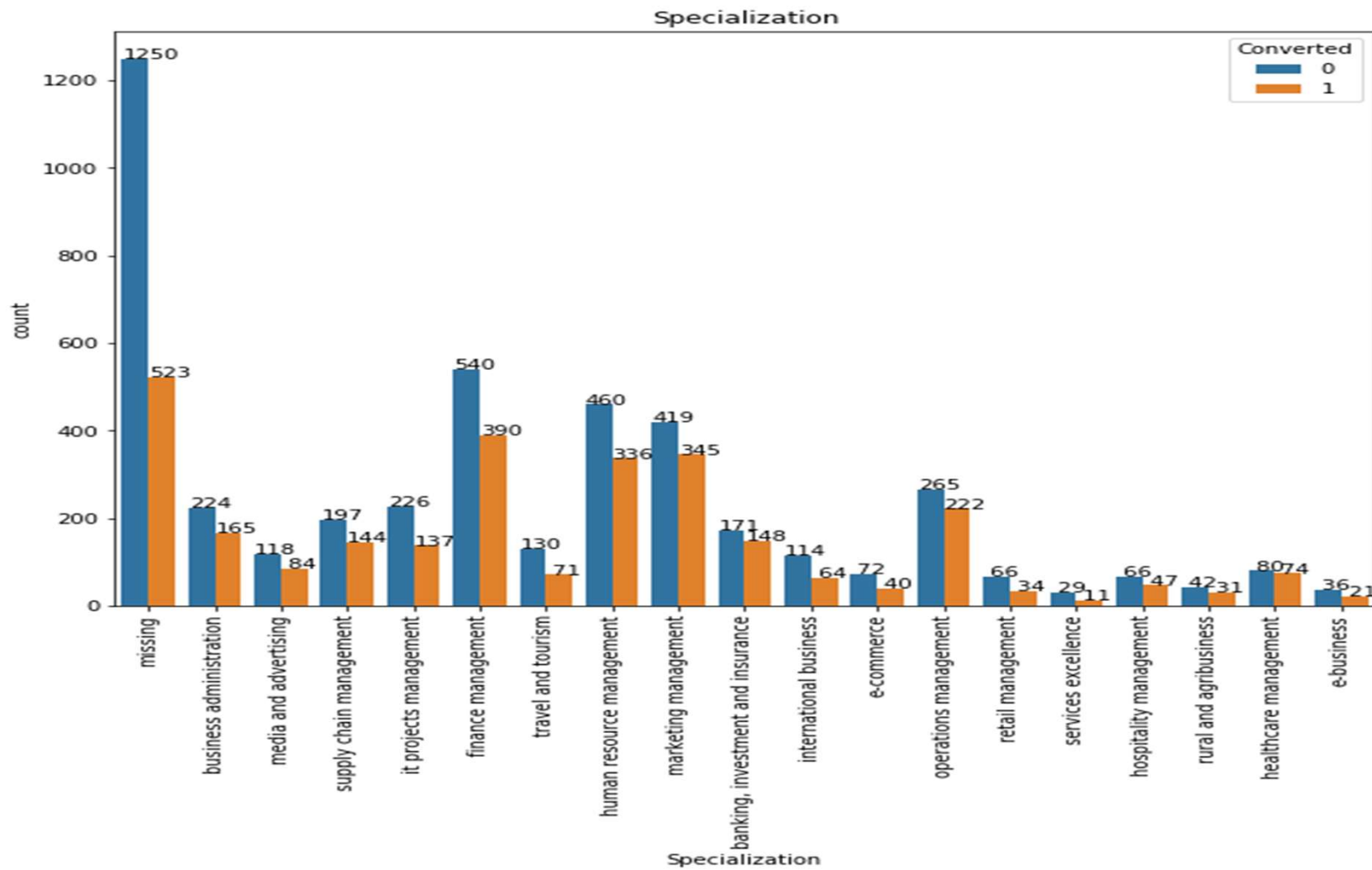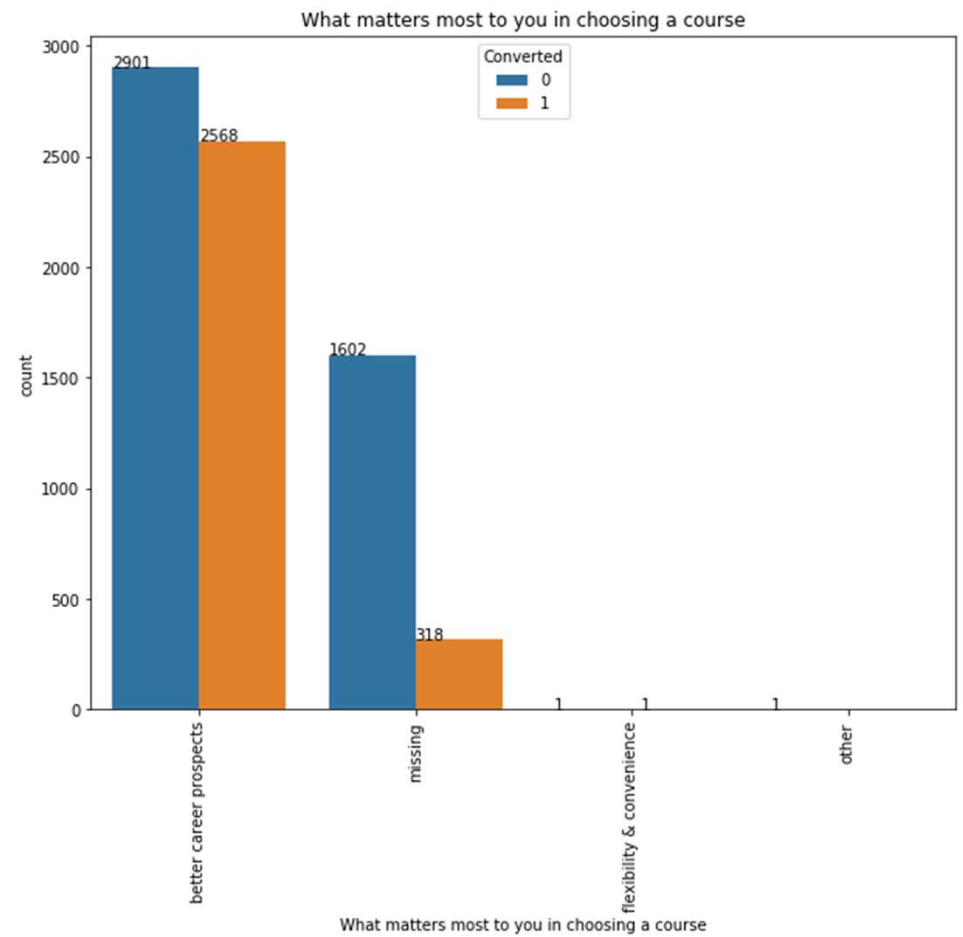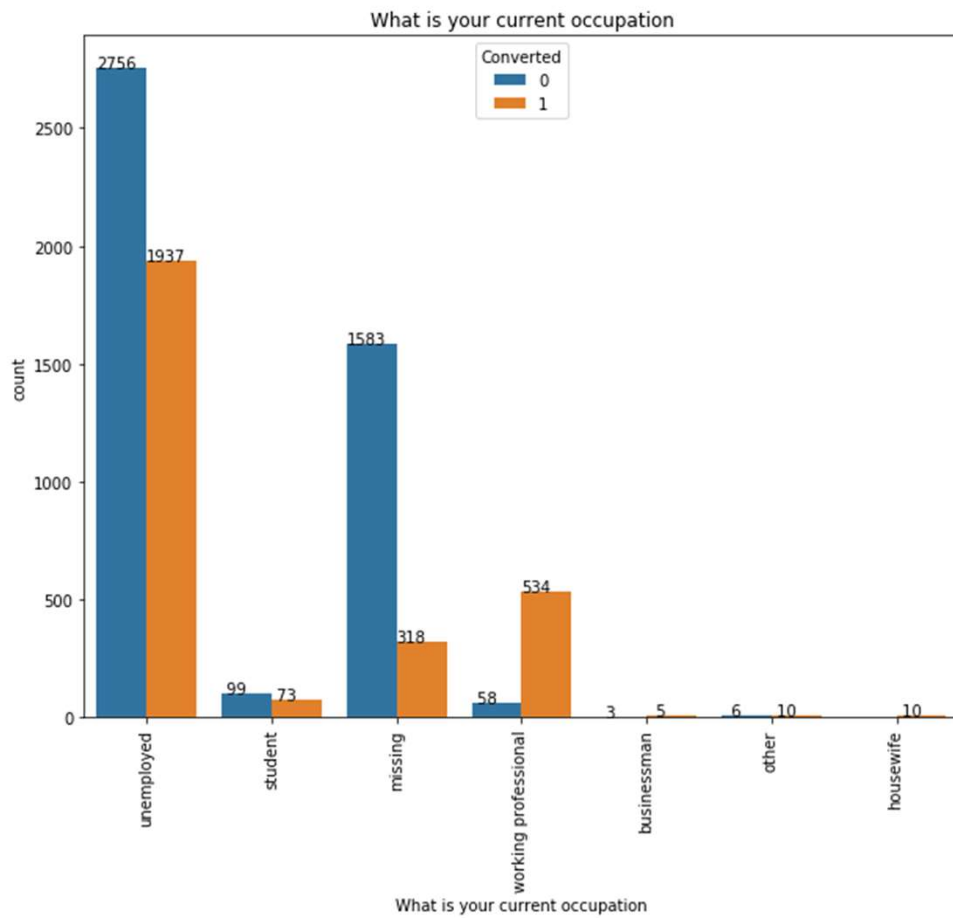
Multivariate Analysis:

Model Building :

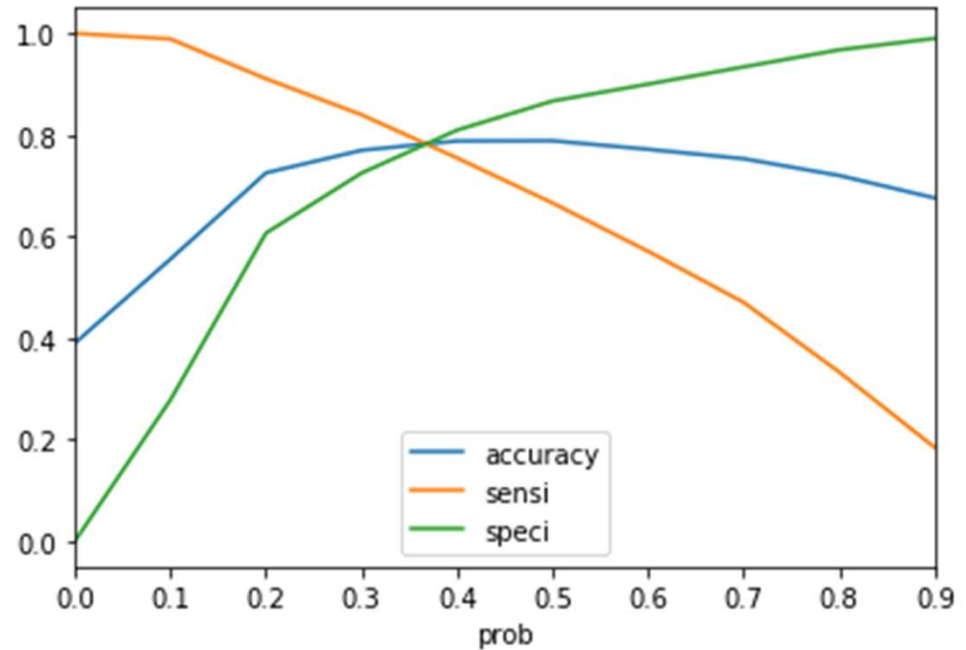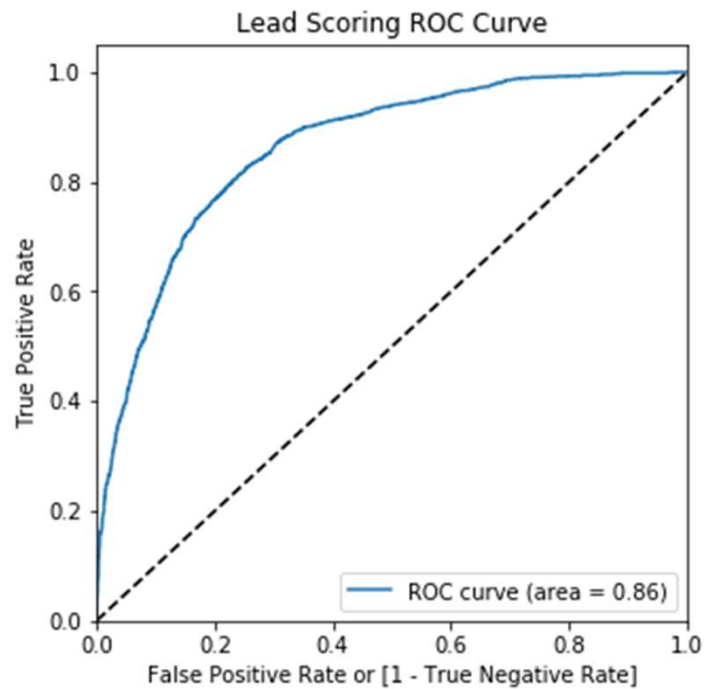➤Splitting the Data into Training and Testing Sets

➤The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio

➤Use RFE for Feature Selection

➤Running RFE with 15 variables as output

➤Predictions on test data set

➤Overall accuracy 80%

Optimize Cut Off (ROC Curve) :

Conclusion :

It was found that the variables that mattered the most in the potential leads are:

➤ The total time spend on the Website.

➤ Page Views Per Visit.

➤ Total Visits

➤When the Last Notable Activity is "Had a phone conversation"

➤When the last Activity is " SMS Sent"

➤When the Load Origin is Load add Format

➤When the Leads Current Profession is "Working Professionals".