LEAD CASE STUDY

NAME: SHAILA
BATCH: DSC-55
Email ID- shailasingal@gmail.com

OBJECTIVE

- 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.

APPROACH

- Identify the missing data
- Identify if there is data imbalance in target column
- To perform Univariate, bivariate, multivariate analysis
- Identify whether the company should give loan to the applicant or not, with specific reasons.
- Identify whether the applicant is genuine or fraud(i.e. not able to pay the loan in previous years)

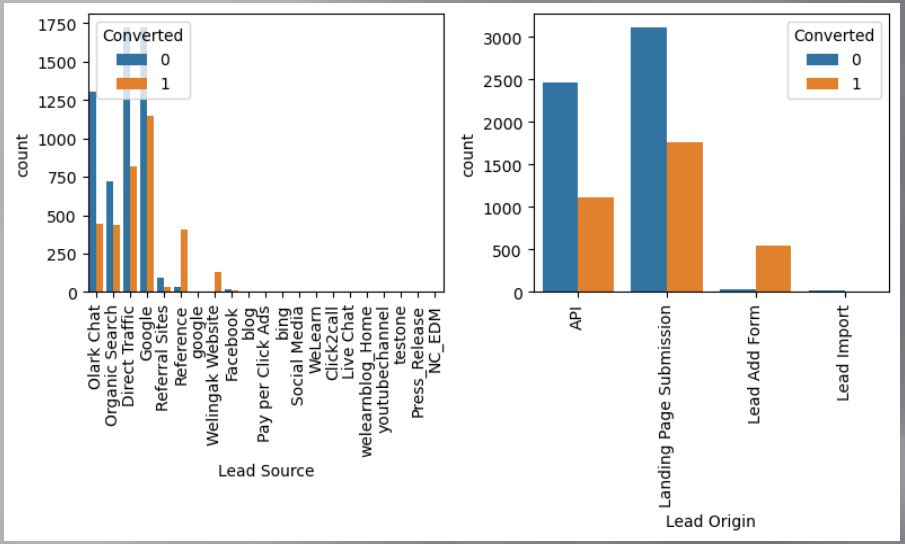
DATASET USED

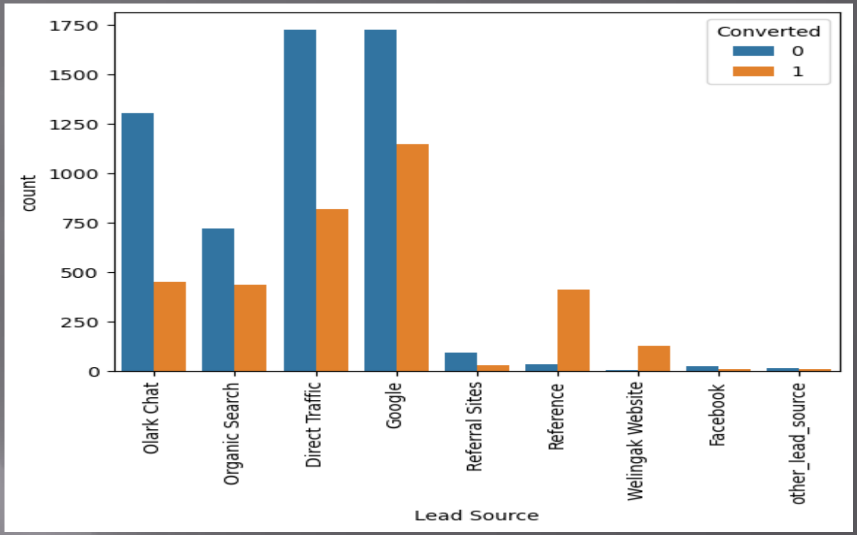
- Leads.csv

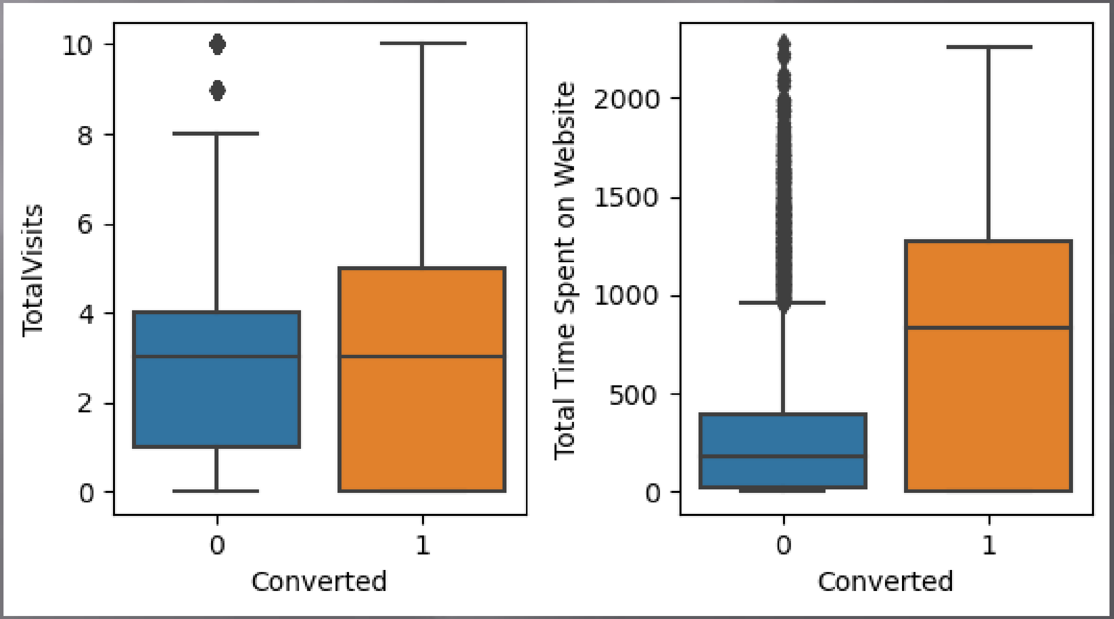
APPROACH AND METHODOLOGY

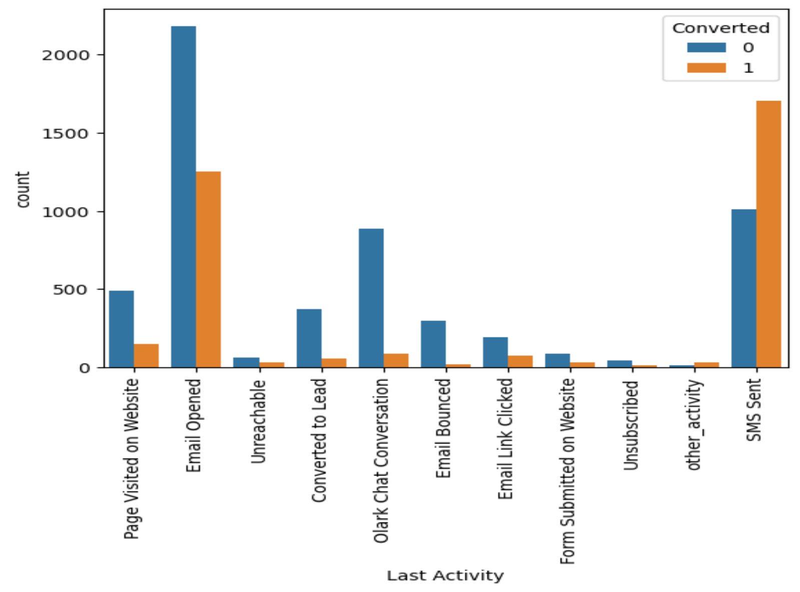
- First step is to identify the null values and impute/remove the null values according to the affect of the Variable on the dataset.
- If there are more than 30-40% null values then we will simply drop the rows. If the null values are less then we will impute the values according to the type of the variable(i.e. if categorical then impute with the mode, if numerical impute with the mean.
- Identify the outliers using boxplot.
- Identifying the data imbalance using countplot.
- EDA (Exploratory Data Analysis)
- Dummy variable creation.
- Train Test Split.
- Model building.
- Feature Selection.
- Model Evaluation.
- Plotting ROC Curve.

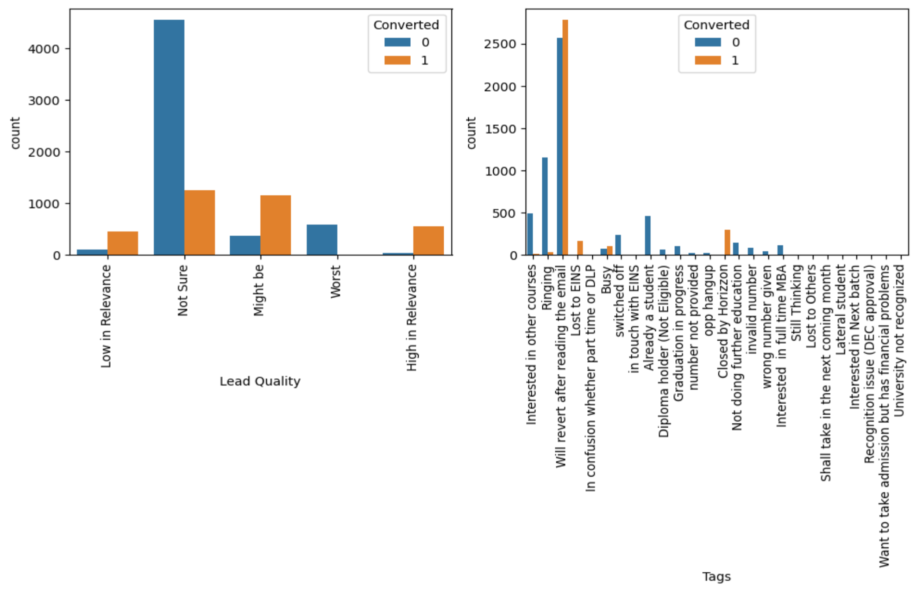
EDA RESULTS

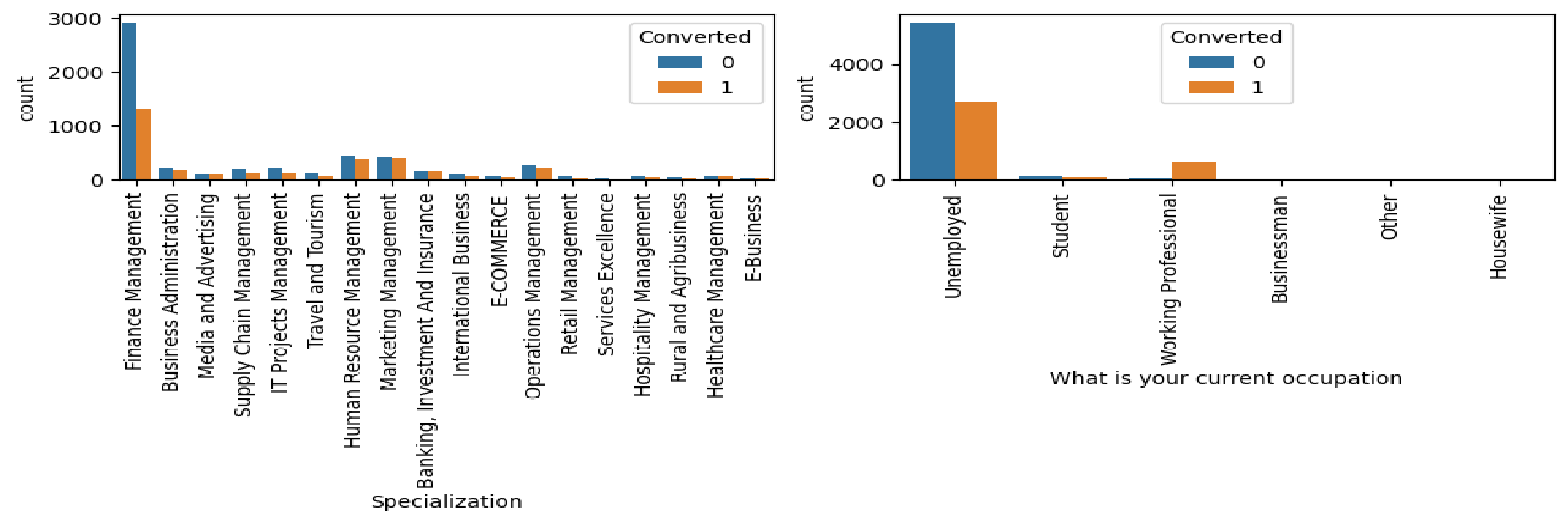

count
1750
1500
1250
1000
750
500
250
0
Converted
0
1
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
google
Welingak Website
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM
Lead Source
count
3000
2500
2000
1500
1000
500
0
Converted
0
1
API
Landing Page Submission
Lead Add Form
Lead Import
Lead Origin


count
1750
1500
1250
1000
750
500
250
0
Converted
0
1
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
Welingak Website
Facebook
other_lead_source
Lead Source


TotalVisits
10
8
6
4
2
0
0
1
Converted
Total Time Spent on Website
2000
1500
1000
500
0
0
1
Converted


Converted
0
1
count
0
500
1000
1500
2000
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
other_activity
SMS Sent
Last Activity


Converted
0
1
count
0
1000
2000
3000
4000
Low in Relevance
Not Sure
Might be
Worst
High in Relevance
Lead Quality
Converted
0
1
count
0
500
1000
1500
2000
2500
Interested in other courses
Ringing
Will revert after reading the email
Lost to EINS
In confusion whether part time or DLP
Busy
switched off
in touch with EINS
Already a student
Diploma holder (Not Eligible)
Graduation in progress
number not provided
opp hangup
Closed by Horizzon
Not doing further education
invalid number
wrong number given
Interested in full time MBA
Still Thinking
Lost to Others
Shall take in the next coming month
Lateral student
Interested in Next batch
Recognition issue (DEC approval)
Want to take admission but has financial problems
University not recognized
Tags

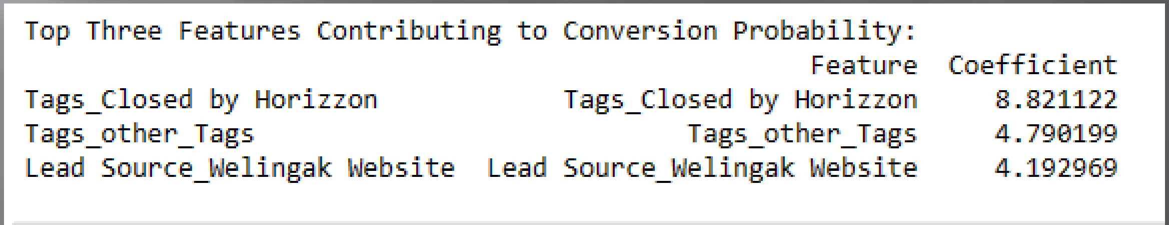

Converted
0
1
count
0
1000
2000
3000
Finance Management
Business Administration
Media and Advertising
Supply Chain Management
IT Projects Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
International Business
E-COMMERCE
Operations Management
Retail Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
Specialization
Converted
0
1
count
0
2000
4000
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation

INSIGHTS FROM EDA ANALYSIS

- ‘API’ and ‘Landing Page Submission’ has less conversion rate but count of leads from them are considerable.
- The count of leads from ‘Lead Add Form’ is low but conversion is very high.
- ‘Lead Import’ has very less count as well as conversion rate hence can be ignored.
- The count of lead from ‘Google’ and ‘Direct Traffic’ is maximum.
- The conversion rate of lead from ‘Reference’ and ‘Welingak Website’ is maximum.
- The median of both conversion and non conversion are same and hence nothing conclusive can be said.
- The count of last activity as 'Email Opened' is maximum.
- The conversion rate of ‘SMS sent as last activity’ is maximum.
- ‘Working professional’ have more conversion rate. ‘Number of unemployed leads’ are more than any other category.
- 'Will revert after reading the email' and 'Closed by Horizon' have high conversion rate.

INSIGHTS FROM LOGISTIC REGRESSION MODEL

- The logistic regression model predicts the probability of the Target variable having a certain value, rather than predicting the value of the target variable directly. Then a cutoff of probability is used to obtain the predicted value of the target variable.
- Here the logistic regression model is used to predict the probability of conversion of a customer.
- Final Logistic Regression model is built with 14 features.
- Top three features contributing in lead conversion are-



```
Top Three Features Contributing to Conversion Probability:
                                                 Feature  Coefficient
Tags_Closed by Horizzon            Tags_Closed by Horizzon     8.821122
Tags_other_Tags                            Tags_other_Tags     4.790199
Lead Source_Welingak Website  Lead Source_Welingak Website     4.192969
```