Lead Score Case Study

Team Details –
Sayli Kulkarni
Vishal Kulthia

Problem Statement

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. 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. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

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 most likely to convert whereas a lower score would mean that the lead 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.

Analysis Approach (1/3)

Step 1. Data Analysis & Preparation

Developing data understanding and correlating it with business and problem statement

Step 2. Exploratory Data Analysis (EDA)

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam a convallis placerat.

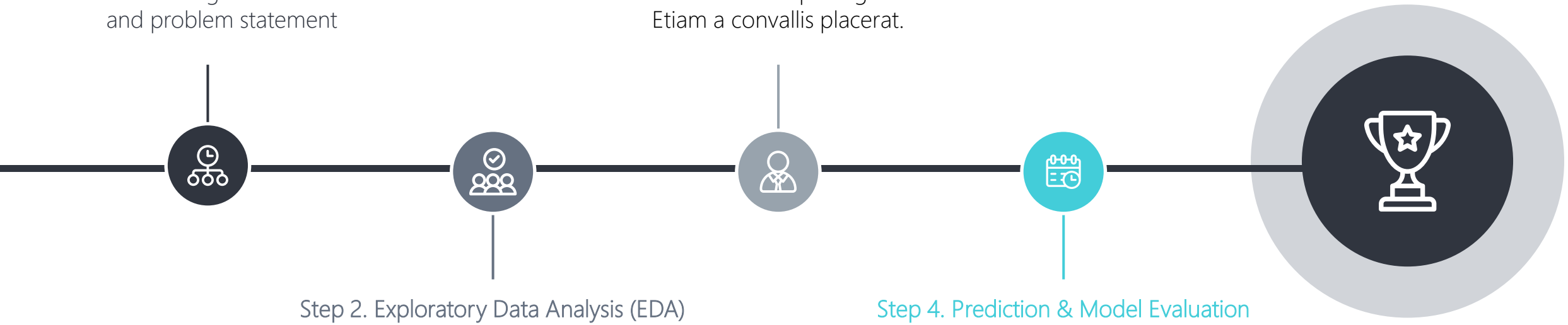
Step 3. Model Building and Performance Analysis

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam a convallis placerat.

Step 4. Prediction & Model Evaluation

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam a convallis placerat.

Final Model & Summary



Your logo

Analysis Approach (2/3)

The below approach was adopted in order to solve the business problem –

Data Preparation & Analysis

1. The data was examined thoroughly by understanding the attributes present, significance of each attribute & data present in them
2. Data quality was examined to figure out presence of outliers, missing values, imbalanced variables & highly correlated variables etc. The same were fixed keeping in mind both business understanding and static tics principles
3. Using this analysis, the redundant columns were dropped and dummy columns were derived
4. The numerical columns were also scaled so as to eliminate any overpowering effect on model prediction due to any column

Exploratory Data Analysis (EDA)

1. On the basis of this analysis various graphs/plots e.g. bar plots, box plots, heat maps etc. are plotted and useful insights were drawn
2. Distribution of categorical variables, univariant & bi variant analysis and corelation between attributes were analysed and same insights were used wherein a business context was required

Analysis Approach (3/3)

Model Building and Performance Analysis

1. Once the attributes were finalized, the data was sliced between train and test data sets. Train data was primarily used to tune the model and decide best attributes for which the performance is optimal
2. Multiple models were developed and few insignificant columns were dropped based on VIF and P value. This not only improves the model performance significantly but also the overall model interpretability becomes easier

Prediction & Model Evaluation

1. Once the attributes & model is finalized then the model performance is evaluated on test data set
2. Metric such as accuracy, precision, recall, specificity, sensitivity etc. were calculated and analysed
3. A ROC curve was also plotted to determine a optimal point/threshold trade-off between specificity & sensitivity is found and the overall model accuracy improves
4. Based on this optimal point the model is again evaluated and metrices mentioned above were calculated
5. Finally the conclusions & recommendations were summarized

Distribution Analysis

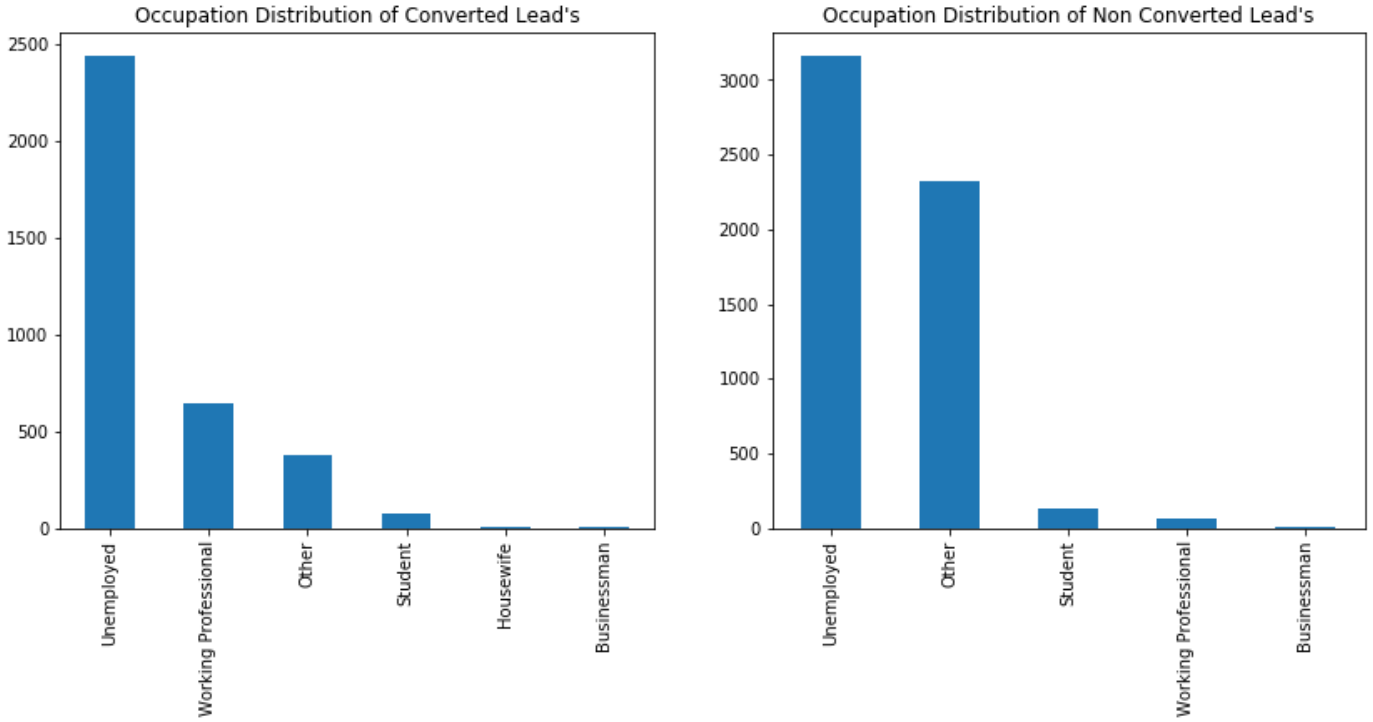


Fig1. Distribution of Occupation of Leads

- In the given data set Category 'Unemployed' has the highest count followed by 'Working Professional' for column occupation's in Converted Lead's. A simple reason could be that unemployed people wants to upskill themselves hence search/opt for such courses more often

- In the given data set Category 'Unemployed' has the highest count followed by 'Student' for column occupation's in Non -Converted Lead's

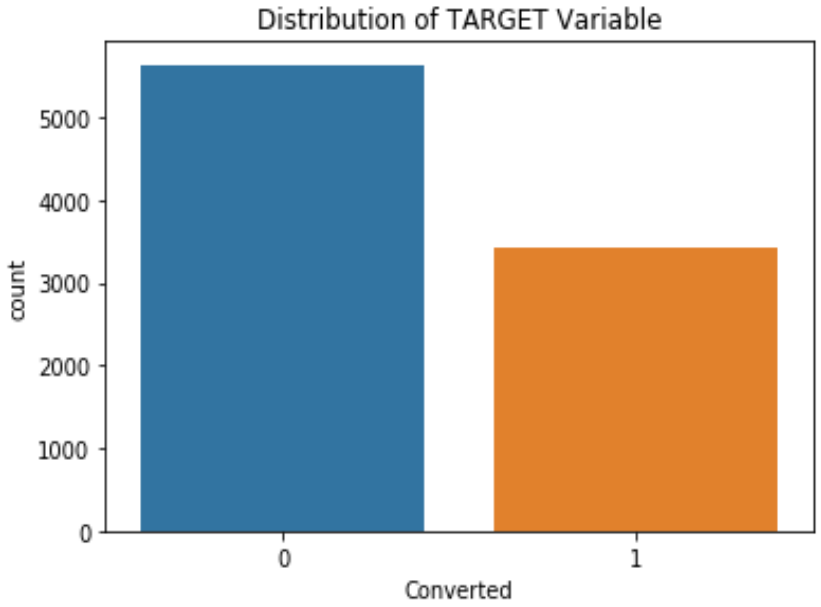


Fig2. Distribution of Target Variables

The distribution shows that in out of ~9K profiles, ~3.3 are customers are have opted for X Education services i.e. leads whereas ~5.8K customers who have not converted/not a lead

Outlier Analysis

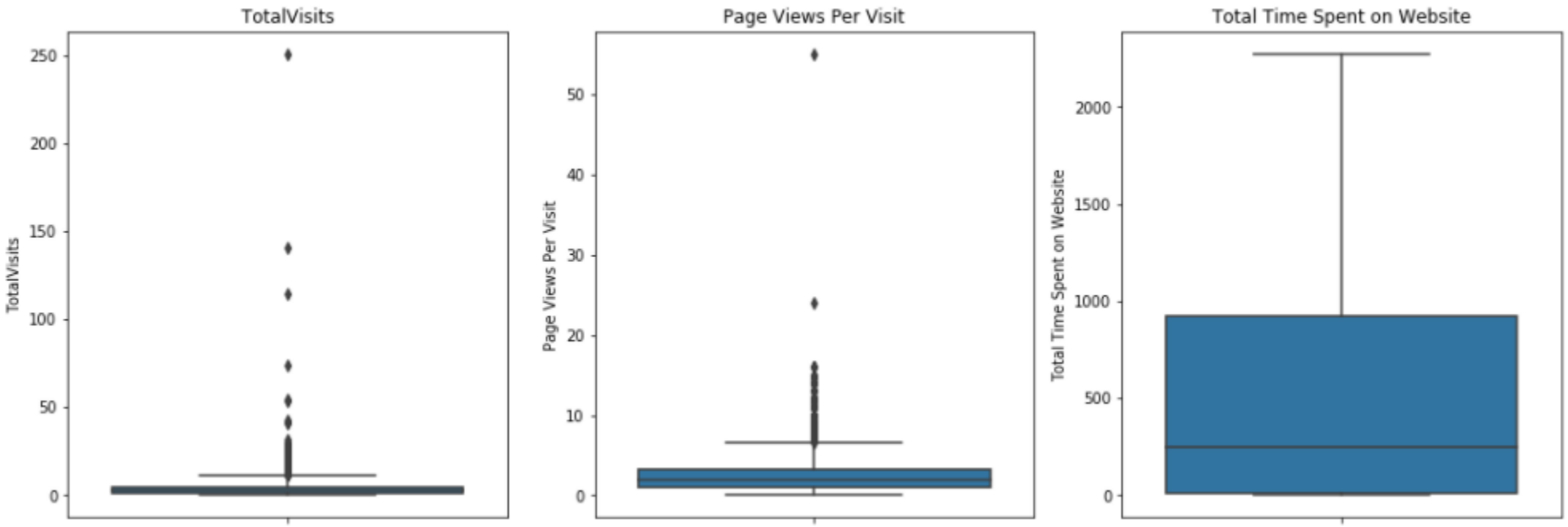


Fig 1. Plot for Total Visit's, Page views per visit and total time spent

1. Outliers are not present in the Total_time_spent_on_website column.
2. Outliers are present in Total_visits and page_views_per_visit column.
3. We will cap these outliers

Univariant Analysis of Numerical Attributes

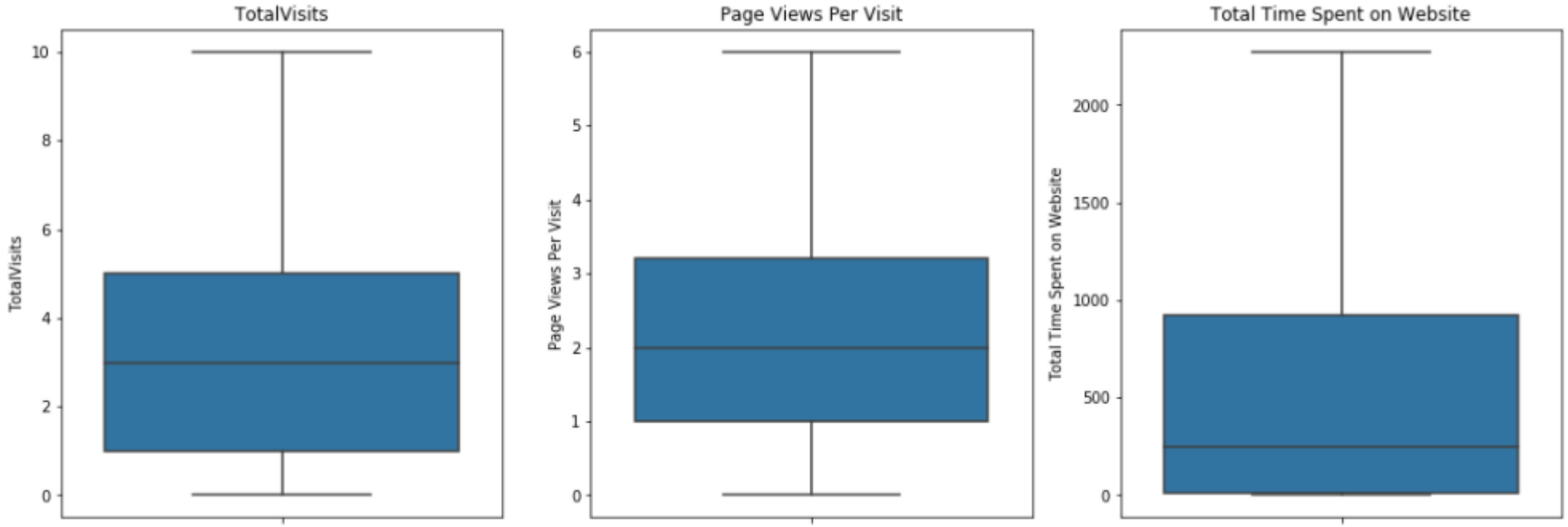


Fig 1. Plot for Total Visit's, Page views per visit and total time spent

1. Quartile for 'TotalVisits' column is present at 3 views
2. In 'Page_views_per_visit' column, median is present at 2 views

Univariant Analysis of Categorical Variables

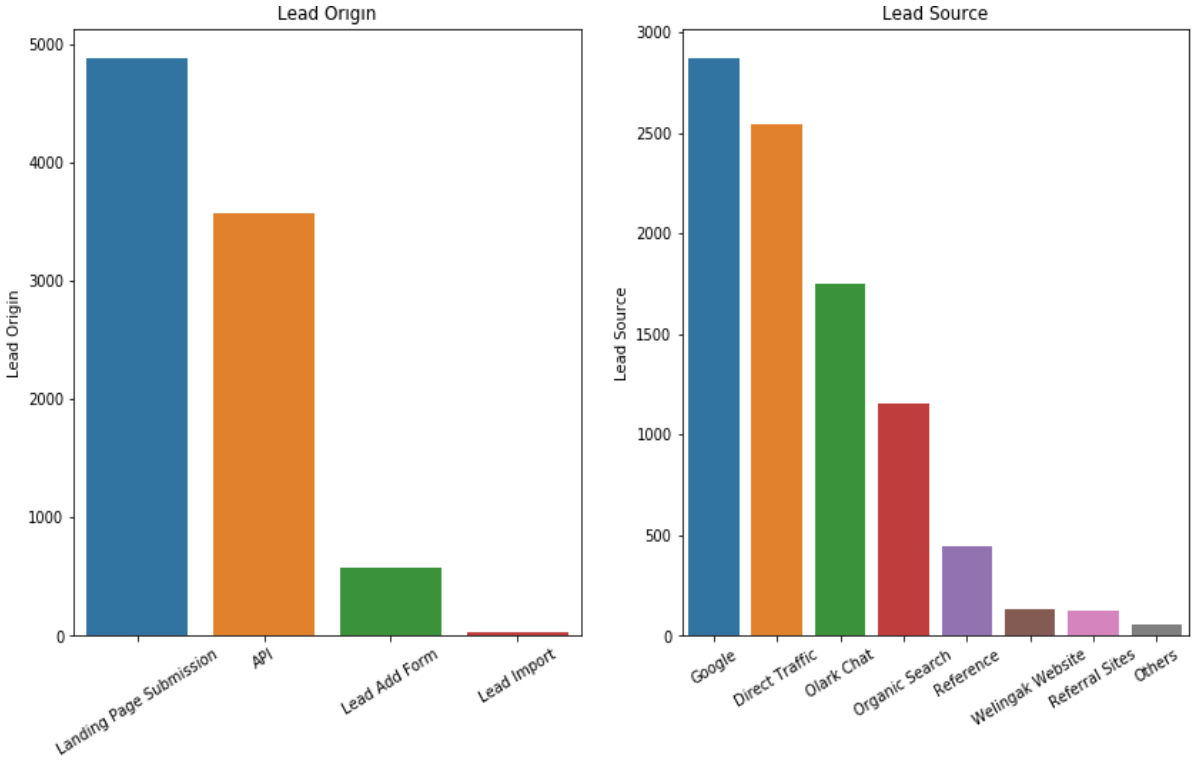


Fig 1. Lead Origin & Lead Score Distribution

The highest count is of category Landing page submission in lead origin and Google in Lead Source

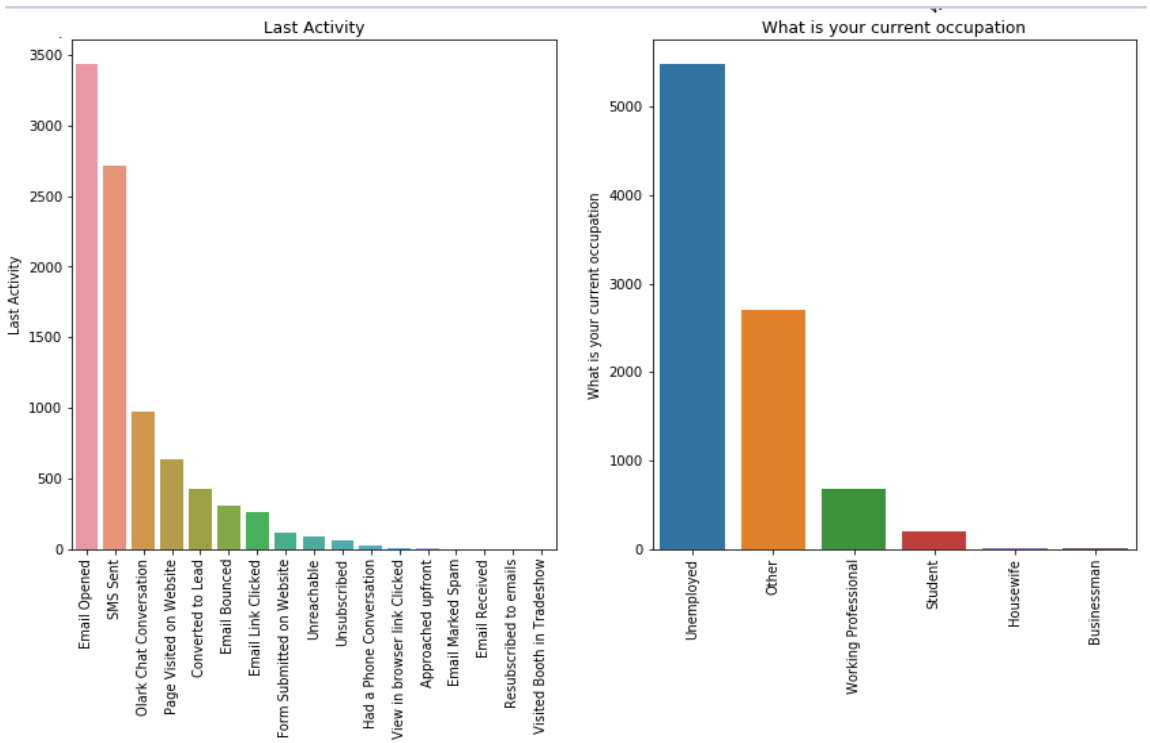


Fig 2. Last Activity vs Current Occupation

The highest count is of category Email Opened in last activity and is Unemployed in case of Occupation

Univariant Analysis

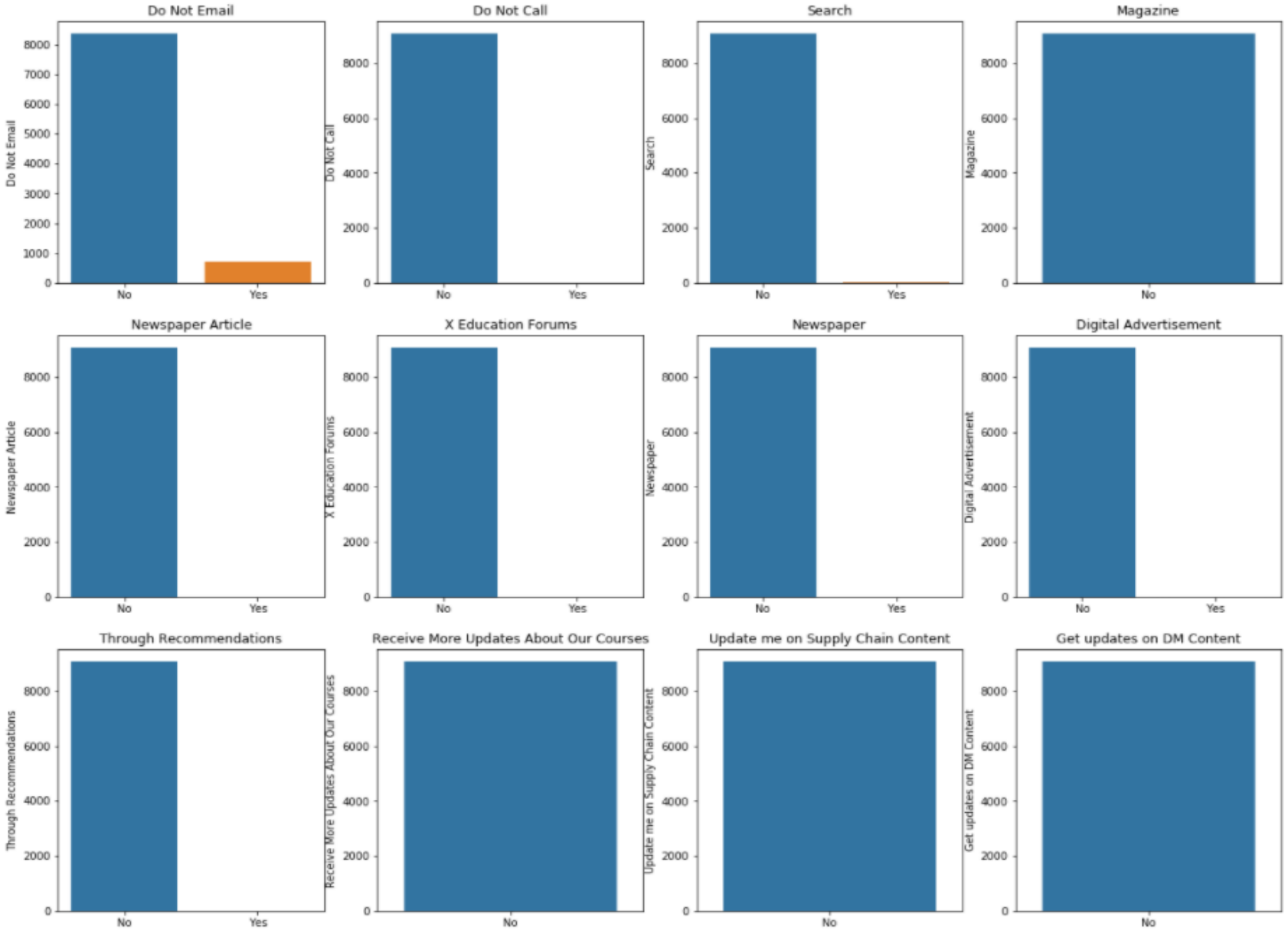


Fig 1. Analysis of Categorical Variables

It can be observed that most of the variables are highly imbalanced in nature thus they are insignificant and can be dropped in model building process

Additionally, Few columns has just one category present e.g. Magazine, Get Updates on DM Content, Update me on Supply Chain Content etc. such column are purely insignificant and doesn't contributes to target variables in any manner

Bivariant Analysis between Numerical- Numerical variables(1/2)

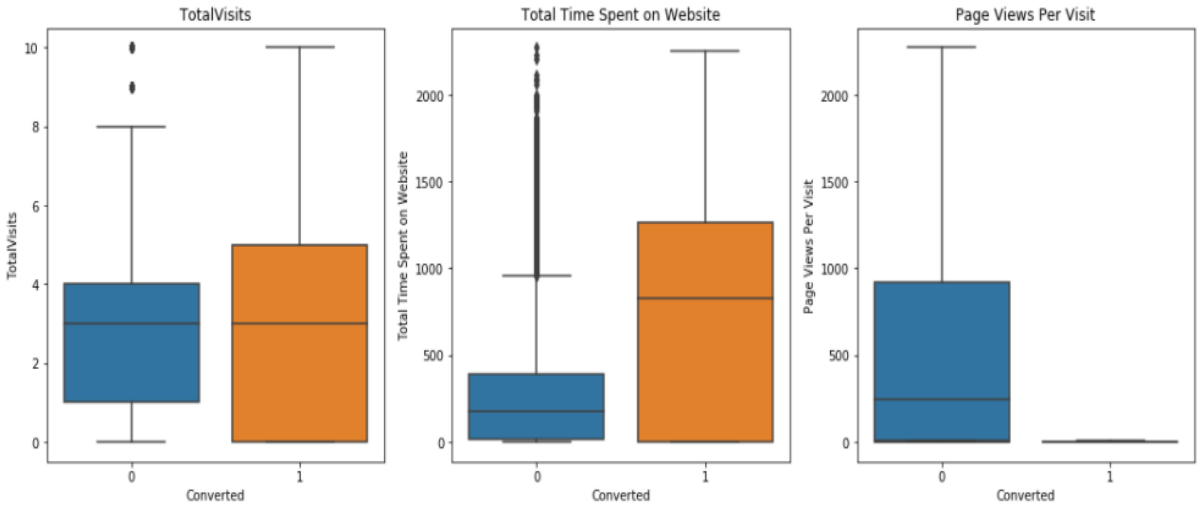


Fig1. Count Plot of Numerical Variables

Quartile for column 'TotalVisits' is same for both leads converted and leads not converted

Median for column 'Total Time Spend on Website' is more leads converted than leads not converted

Data is not present for leads converted in Page views per visit column

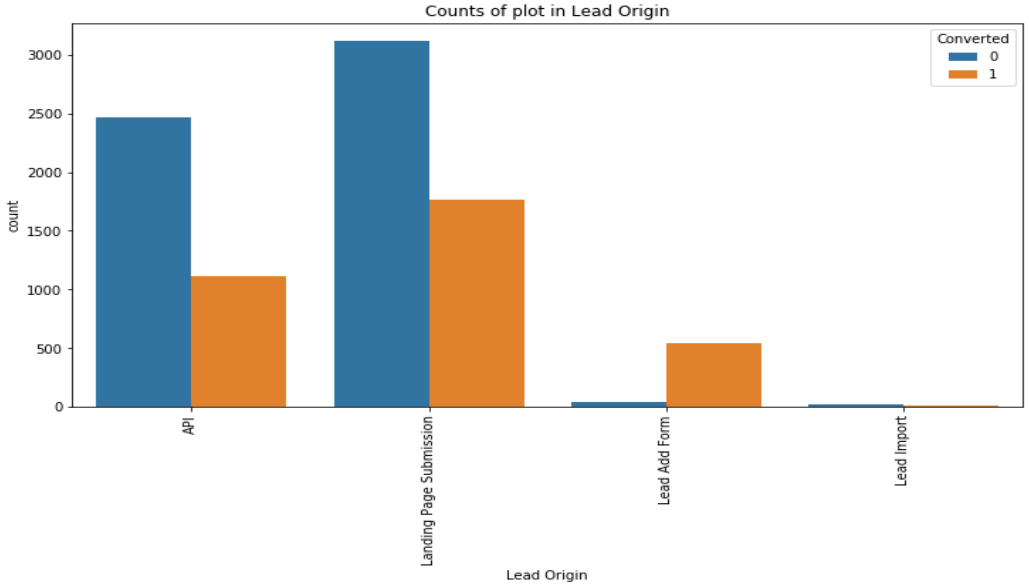


Fig 2. Count Plot of Lead Origin

Count of leads converted is less than leads not converted for every category

Most of leads are getting converted in landing page submission category as compared to others

For 'Land Add Form' count of leads converted is more than not converted

Bivariant Analysis on Numerical Variables (2/2)

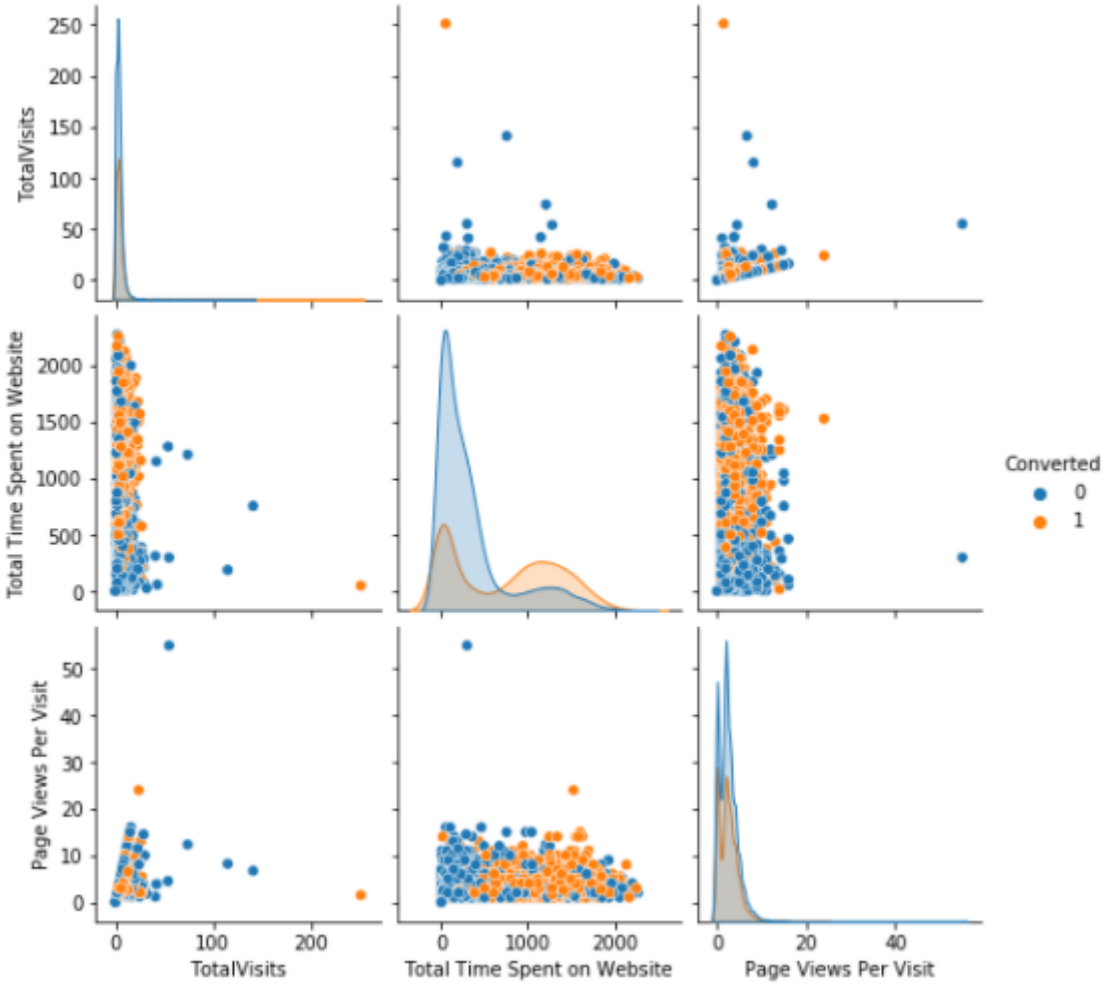


Fig1. Pair Plot Of Categorical Variables

The plot shows corelation pattern between various numerical variables

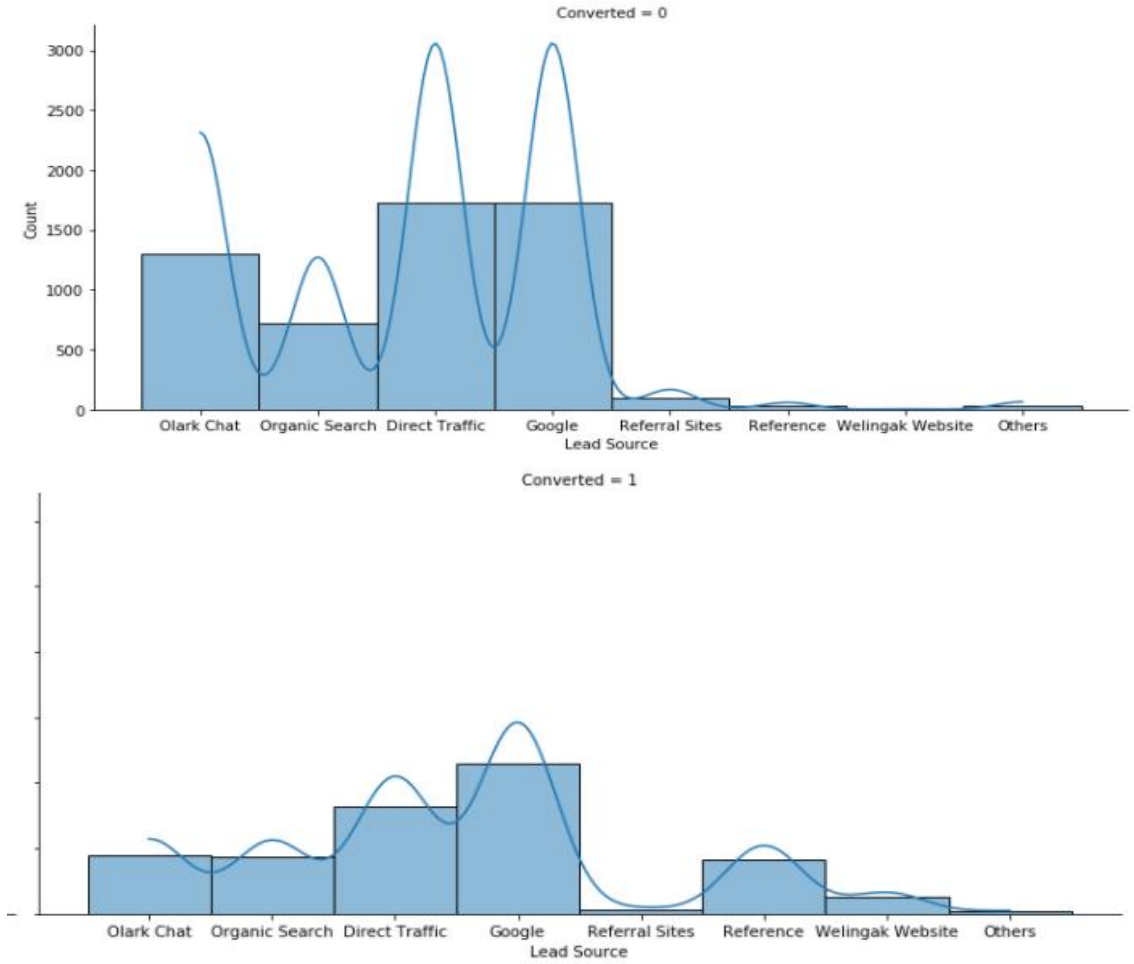


Fig 2. Pair Plot Of Categorical Variables

The plot shows corelation between Lead Source and Target Variable (Both for Converted =1 & Converted =0)

Correlation Analysis (1/2)

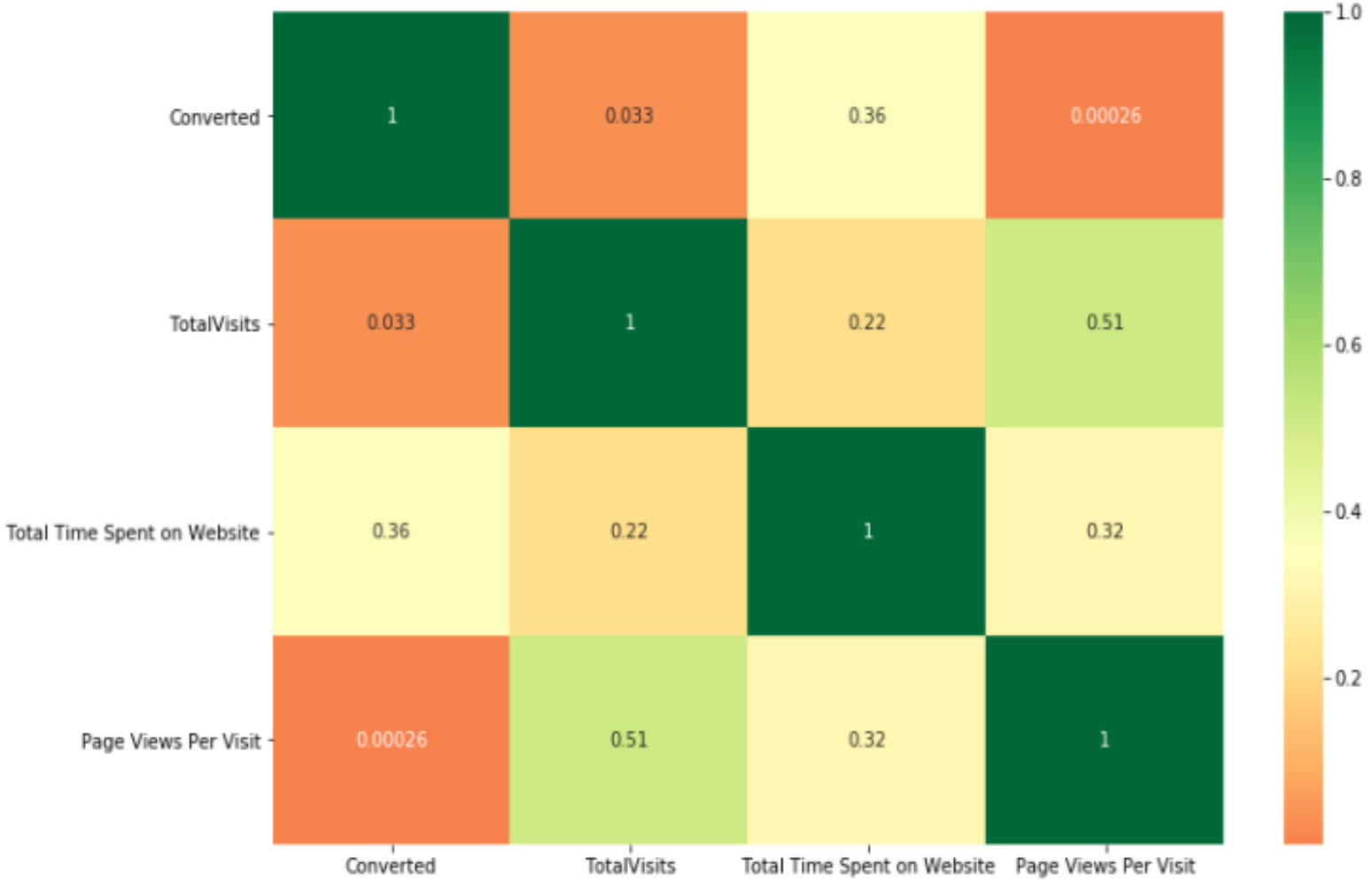


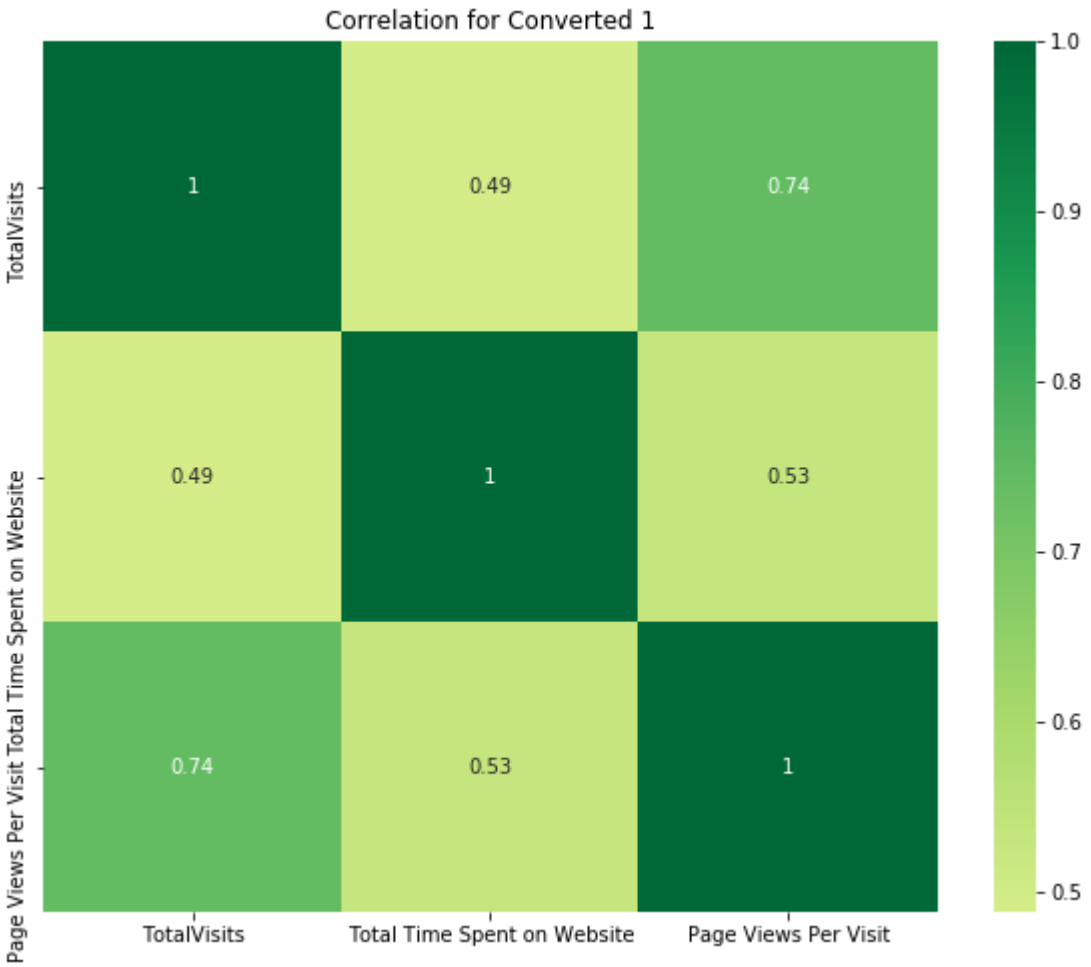
Fig 1. Heat Map – Correlation Analysis

A heat map has been plotted to visualize the correlation between variables. I.e. understanding characteristic of variable and effect on it due to other attribute

It can be observed that Total Visits and Page View Per Visit has a correlation of 0.51.

Additionally Total Time Spent has a correlation of 0.36 with target variable (converted). This also make sense as a customer is more interested therefore time invested by him/her in researching about product on website is high thereby high correlation value

Correlation Analysis (2/2)



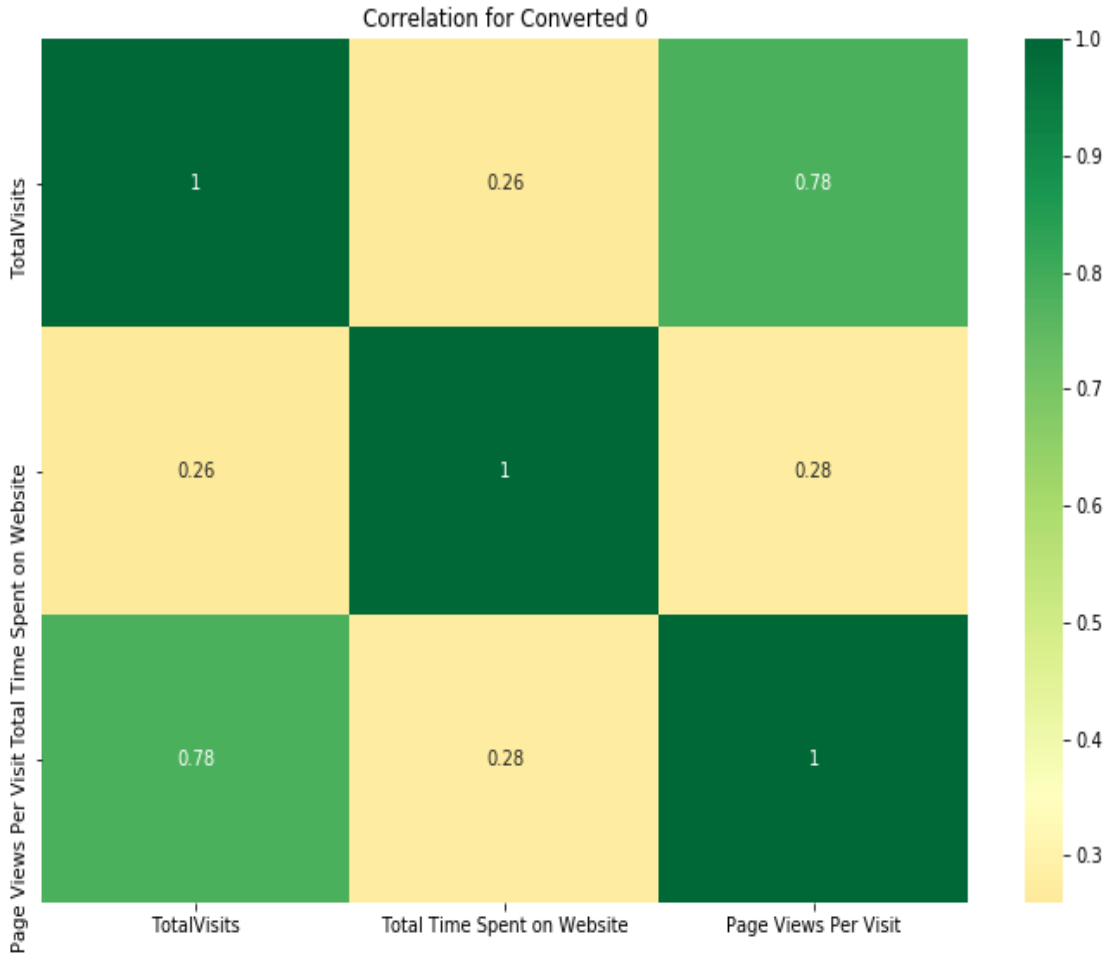
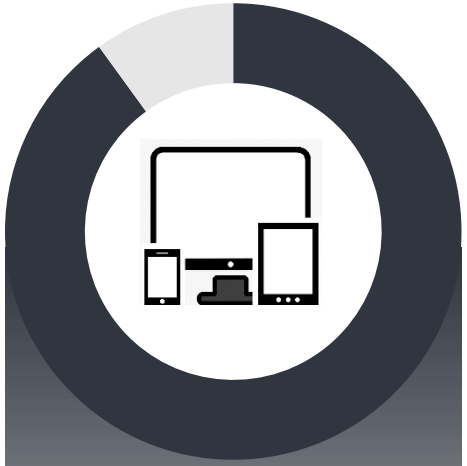


Fig1. Heat map for Correlation for Converted 1
Strong correlation present between page views per visit to total visit

Fig2. Heat map for Correlation for Converted 0
There is strong correlation present between TotalVisit and Converted

Analysis Outcome/Inferences

Effective Usage of Social Media Platforms

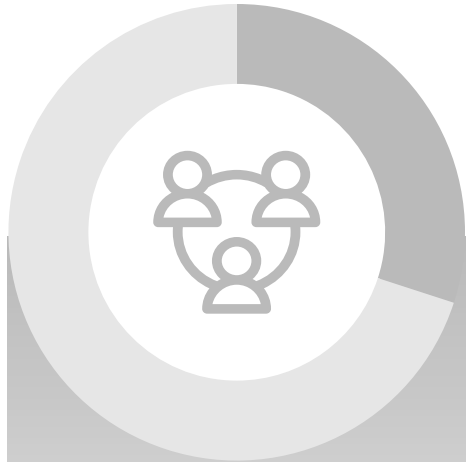


Analysis Outcome & Final Model shows that lead sources obtained from social media platform such as Welingak Website, Olark Chat positively affects the probability of lead conversion

Prioritizing Customer Occupation

It has been observed that Occupation is a very crucial criteria to increase the lead conversion. Analysis depicts that Unemployed & Working Professional is the best occupation that X Education shall focus

Time Spent on Website



It has been observed that customers who spend appreciable time on X Education Website have high chances of conversion to Lead. These customer's shall be treated on priority


THANK YOU