greatlearning

PROJECT REPORT

LEAD SCORING

Batch: DSE_BANGALORE_FEB2020

Submitted by:

Neha Jindal
Aviraj Roy
Lovely Kumari
Mohan CR
Nikhil Niket

Mentor:

Anjana Agrawal

greatlearning

Table of Contents

Industry Review

An education company ‘X Education’ sells online courses to people from different occupations. It markets its courses on several websites and search engines like Google. Once these people land on the website, they browse the courses or fill up a form for the course or watch some demo videos. When these people fill up a form providing their email address or phone number, they are classified to be a *Lead.*

Companies often gather a tremendous amount of data, such as browsing behaviour, email activities, and other contact data. This data can be the source of important competitive advantage by utilizing it in estimating a contact's purchase probability using predictive analytics. When 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. Companies often gather a tremendous amount of data, such as browsing behaviour, email activities, and other contact data.

To increase the efficiency in tracking hot leads we will apply machine learning. Here, machine learning can be used to perform lead scoring as a special application case of purchase probabilities and approaching a possible lead who purchases the product. Historical behavioural data is used as training data for the classification algorithm.

SCOPE: A calculated purchase probability is used by companies to solve different business problems, such as optimizing their sales processes and going about approaching the hot leads. This data can be the source of important competitive advantage by utilizing it in estimating a contact's purchase probability using predictive analytics.

Project Justification

Project Statement: An education company ‘X Education’ sells online courses to people from different occupations. It markets its courses on several websites and search engines like Google. Once these people land on the website, they browse the courses or fill up a form for the course or watch some demo videos. When these people fill up a form providing their email address or phone number, they are classified to be a Lead. When 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%. We aim at increasing this Lead Conversion rate.

Complexity involved: A business might have many potential customers leads but most of those customers will not turn into actual, paying customers in the end. A sales team must sort through a long list of potential customers and figure out how to spend their time. That is where lead scoring comes in. This is a system that analyses attributes about each new lead in relation to the chances of that lead becoming a customer and uses that analysis to score and rank all of the potential customers. With that new ranking, the sales team can then prioritize their time, and only spend time on the leads that are highly likely to become paying customers.

Project Outcome: Companies often gather a tremendous amount of data, such as browsing behaviour, email activities, and other contact data. A Machine Learning Model which gives the calculated purchase probability based on the data available can be used by companies to solve different business problems, such as optimizing their sales processes and going about approaching the hot leads and converting them to a customer. It is a highly valued process in marketing of sales products

Data Dictionary

1. Shape: The dataset consists of 37 columns with 9,240 rows of data with the target variable 'Converted'.

2. Description: Below is the list of all columns with their description

Variables	Description	Data Types
Prospect ID	A unique ID with which the customer is identified	Numerical - Unique IDs
Lead Number	A lead number assigned to each lead procured.	Numerical - Unique IDs
Lead Origin	The origin identifier with which the customer was identified to be a lead. Includes API, Landing Page Submission, etc.	Categorical
Lead Source	The source of the lead. Includes Google, Organic Search, Olark Chat, etc.	Categorical
Do Not Email	An indicator variable selected by the customer wherein they select whether of not they want to be emailed about the course or not.	Categorical
Do Not Call	An indicator variable selected by the customer wherein they select whether of not they want to be called about the course or not.	Categorical
Converted	**TARGET VARIABLE**. Indicates whether a lead has been successfully converted or not.	Numerical discrete
TotalVisits	The total number of visits made by the customer on the website.	Numerical continuous
Total Time Spent on Website	The total time spent by the customer on the website.	Numerical continuous
Page Views Per Visit	Average number of pages on the website viewed during the visits.	Numerical continuous
Last Activity	Last activity performed by the customer. Includes Email Opened, Olark Chat Conversation, etc.	Categorical
Country	The country of the customer.	Categorical
Specialization	The industry domain in which the customer worked before. Includes the level 'Select Specialization' which means the customer had not selected this option while filling the form.	Categorical
How did you hear about X Education	The source from which the customer heard about X Education.	Categorical

What is your current occupation	Indicates whether the customer is a student, umemployed or employed.	Categorical
What matters most to you in choosing this course	An option selected by the customer indicating what is their main motto behind doing this course.	Categorical
Search	Indicating whether the customer had seen the ad in any of the listed items.	Categorical
Magazine		Categorical
Newspaper Article		Categorical
X Education Forums		Categorical
Newspaper		Categorical
Digital Advertisement		Categorical
Through Recommendations	Indicates whether the customer came in through recommendations.	Categorical
Receive More Updates About Our Courses	Indicates whether the customer chose to receive more updates about the courses.	Categorical
Tags	Tags assigned to customers indicating the current status of the lead.	Categorical
Lead Quality	Indicates the quality of lead based on the data and intuition the the employee who has been assigned to the lead.	Categorical
Update me on Supply Chain Content	Indicates whether the customer wants updates on the Supply Chain Content.	Categorical
Get updates on DM Content	Indicates whether the customer wants updates on the DM Content.	Categorical
Lead Profile	A lead level assigned to each customer based on their profile.	Categorical
City	The city of the customer.	Categorical
Asymmetrique Activity Index	An index and score assigned to each customer based on their activity and their profile	Categorical
Asymmetrique Profile Index		Categorical
Asymmetrique Activity Score		Numerical discrete
Asymmetrique Profile Score		Numerical discrete
I agree to pay the amount through cheque	Indicates whether the customer has agreed to pay the amount through cheque or not.	Categorical
a free copy of Mastering The Interview	Indicates whether the customer wants a free copy of 'Mastering the Interview' or not.	Categorical
Last Notable Activity	The last notable acitivity performed by the student.	Categorical

Variable categorization:

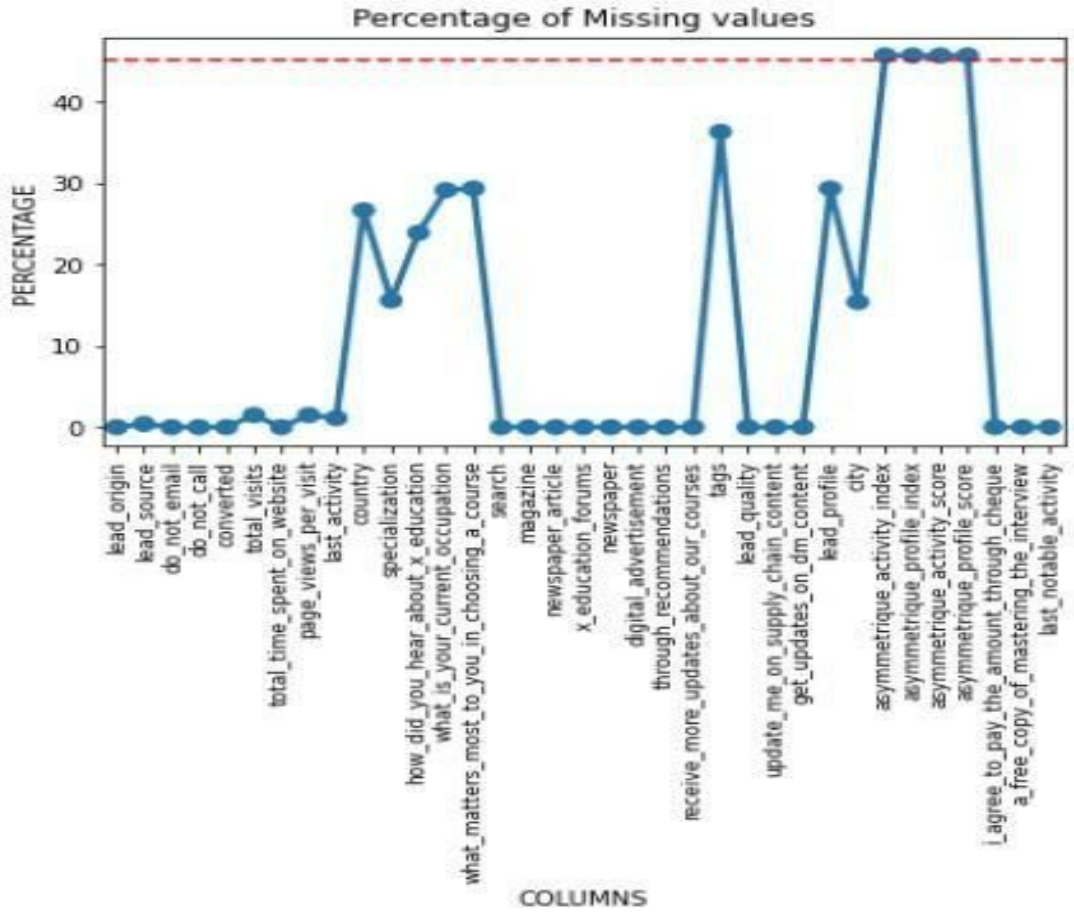
Numerical variables Count = 30
Categorical variables Count = 7

Redundant Columns:

Two variables Lead_number, prospect_id have 100% Unique values

Magazine, receive more updates about our courses, Update me on Supply Chain Content, Get updates on DM content, I agree to pay amount through check: Have 100% same value

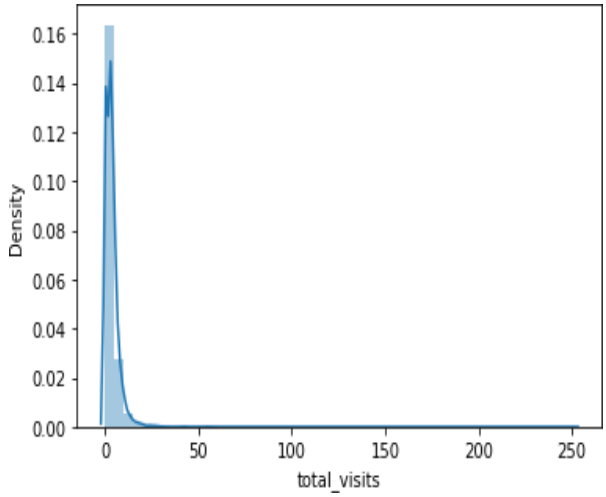
Missing Value Analysis:

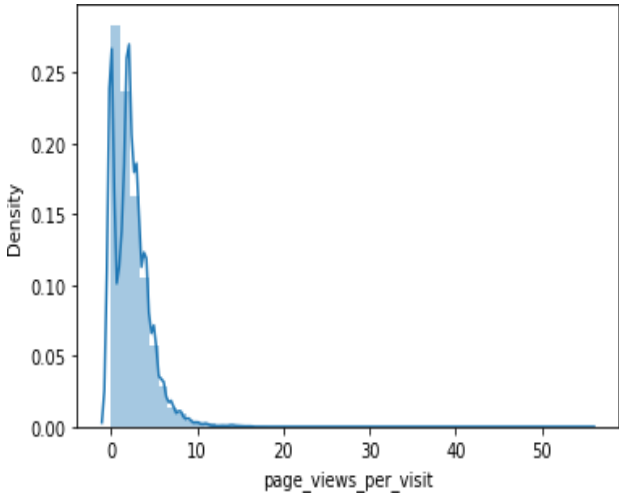


Data Exploration

Null Value Treatment:

Total Visits & Page views per visit: From the distribution plot of both we see that the data is right skewed and if we replace it with mean it would become more skewed. Hence, we replace it with Median.





For categorical features with less percentage of null values, we used **Pattern Matching** with respect to an existing feature. We derived patterns from non-null rows and using those imupted missing values in semi-null rows.

As the count of null values increased, pattern matching did not yield any results, so we used **Method Fill** for null value treatment.

Variables	Null Value Treated by	Null Values in Treatment
TotalVisits	Median	137
Page Views Per Visit	Median	137
Lead Source	Pattern Matching	36
Last Activity	Pattern Matching	103
Country	Method Fill	2461
What is your current occupation	Method Fill	2690
Lead Profile	Method Fill	2709
What matters most to you in choosing this course	Method Fill	2709
Tags	Method Fill	3353

Specialization & How did you hear about X Education & City: Some values in these features were as Select. We converted them to null and used Method Fill to replace all the null values

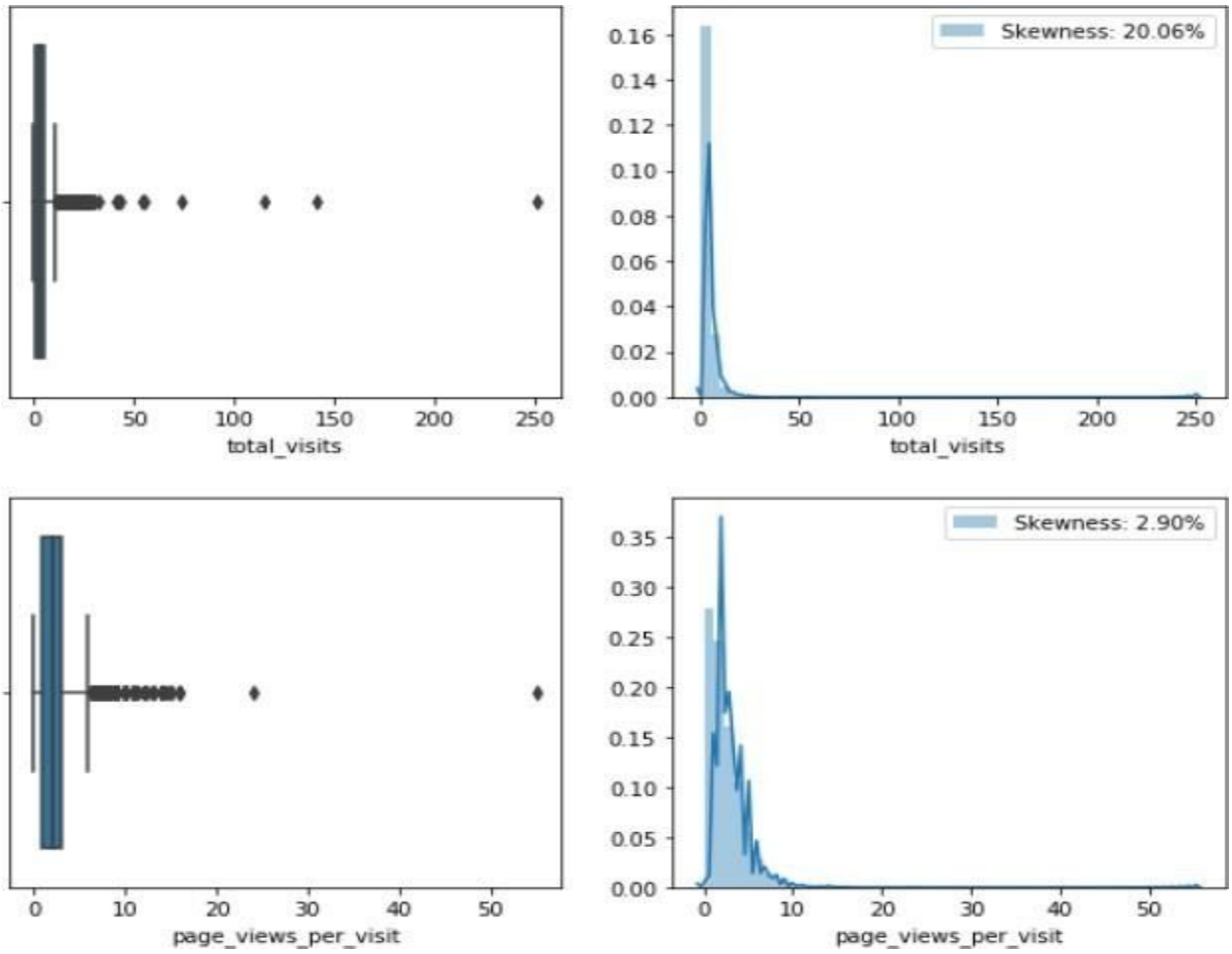
Variables	Null Value Treated by	Null Values in Treatment
Specialization	Assume Select as null and used Method Fill	1438
How did you hear about X Education	Assume Select as null and used Method Fill	2207
City	Assume Select as null and used Method Fill	1420
Lead Quality	Replaced Null with 'Not Sure'	4767

Lead Quality had the highest percentage of null values so we converted it to a new label named 'Not Sure'

Outlier Treatment

We have two numeric variables which have a high number of outliers: total_visits and page_views_per_visit

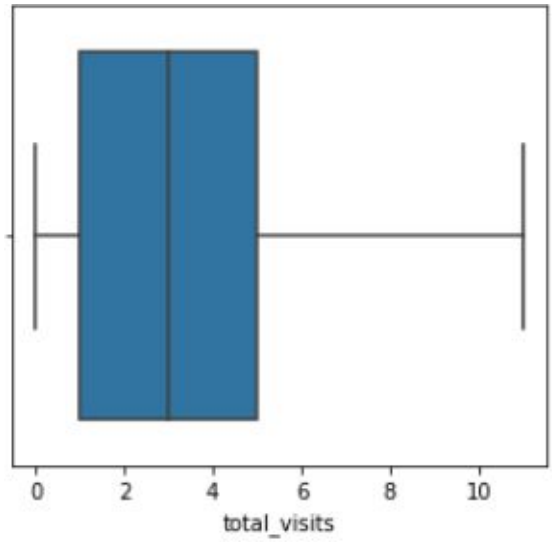
The outliers are treated as per InterQuartile Range technique. Here, any data point less than

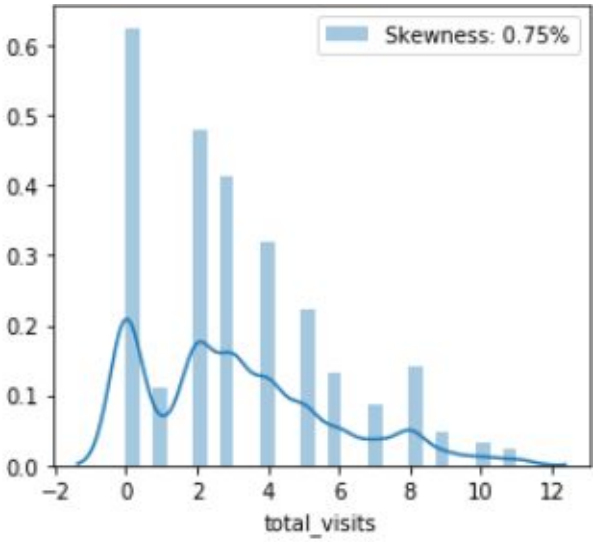


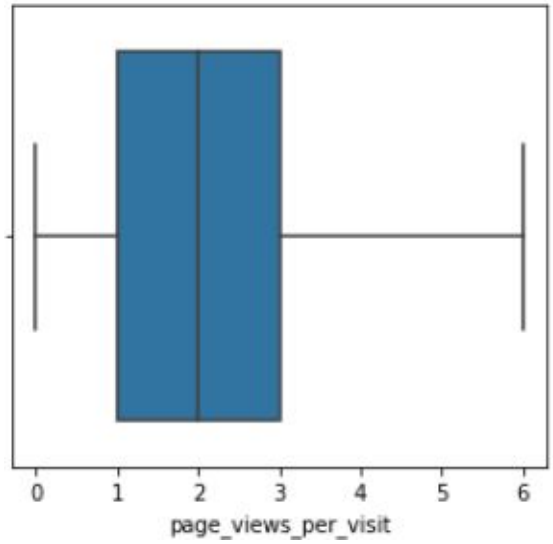
the Lower Bound or more than the Upper Bound is considered as an **outlier**, where Lower Bound is 1.5*IQR below the first quartile and Upper Bound is 1.5*IQR above the third quartile. Now, all outliers will be replaced by 0.01 quantile value for Lower Bound outliers and 0.99 quantile value for Upper Bound outliers.

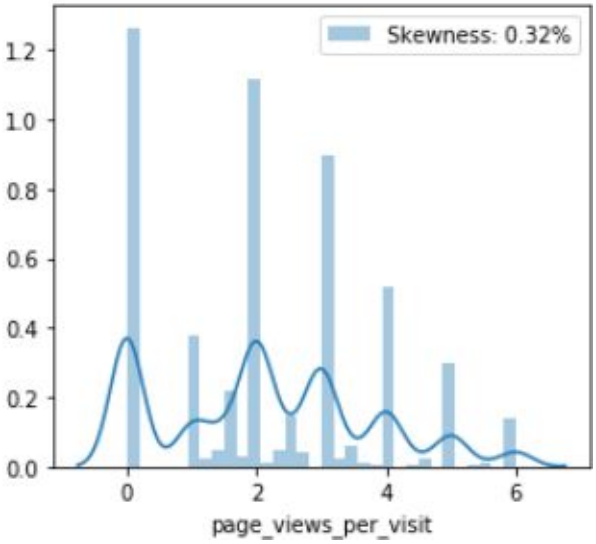
Upon treatment, the variables transform considerably. We can see a difference as the distribution becomes more Gaussian when outliers disappear from the features.

1. Skewness of total_visits has decreased from 20.06 to 0.75
2. Skewness of page_views_per_visit has decreased from 2.9 to 0.32

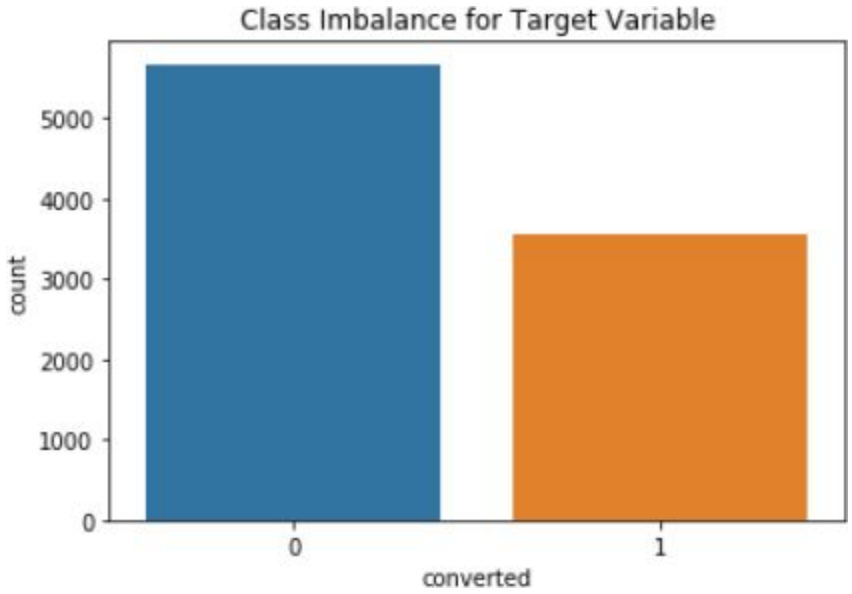








Class Imbalance:

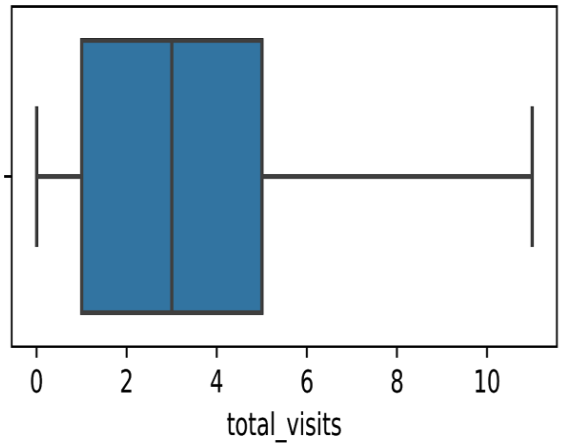


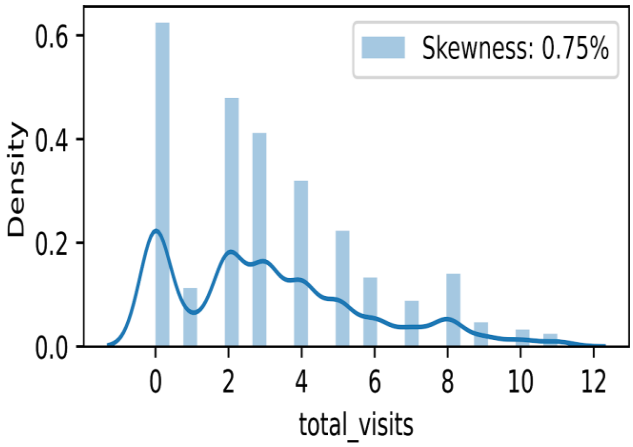
There is no observed class imbalance

Univariate Analysis:

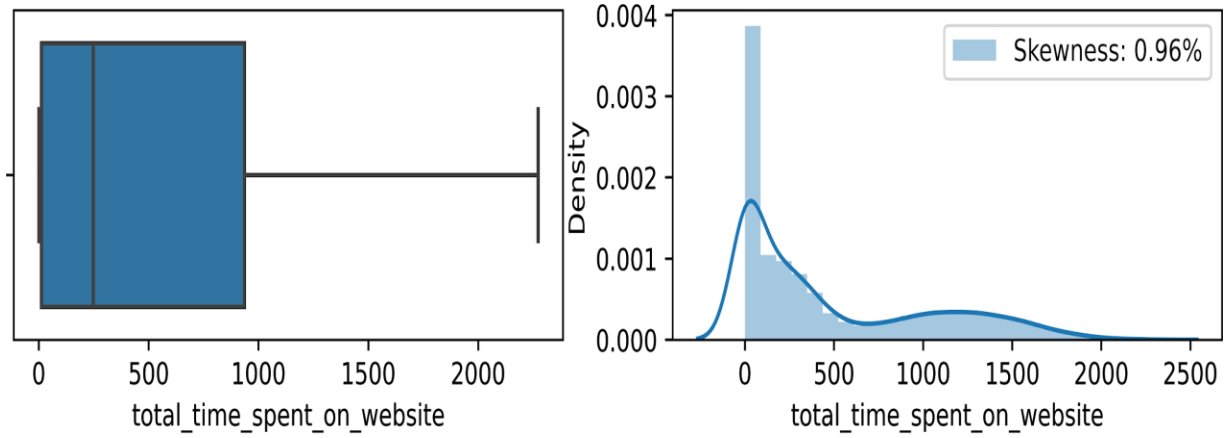
Numerical Columns:

1. TOTAL VISITS: The distribution is slightly right skewed with 0.75 skewness. As per the distribution, less visits come from people who tend to visit the site once and never come back. The customers who want to gain an idea about X education visit 1-5 times. The highest visits come from the most engaged customers, creating the skewness in the distribution.

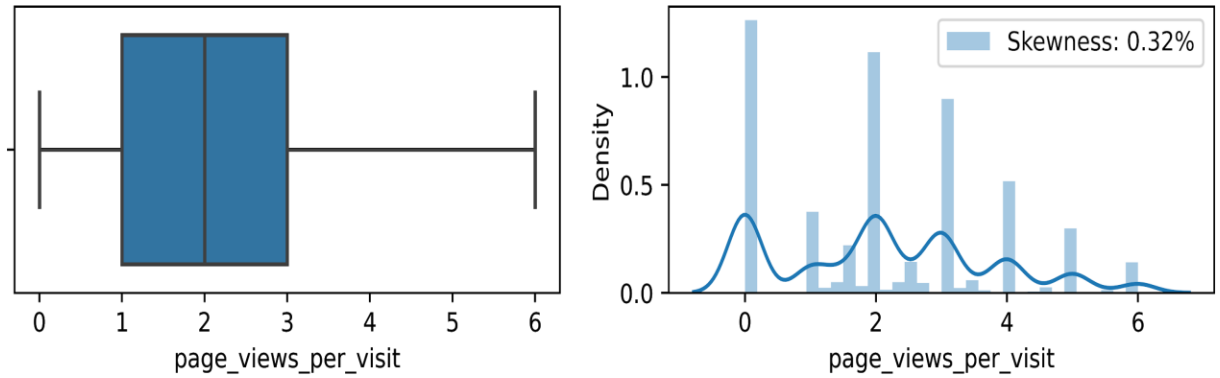




2. TOTAL TIME SPENT ON WEBSITE: The distribution is right skewed with 0.96 skewness. As per the distribution, the least engagement of the website is from leads who tend to spend less time on the website. Quite a few leads spend up to 500 minutes. This comes from leads who want to gain an idea about X education. The people who spend 1000 minutes on the website are the most engaged leads.

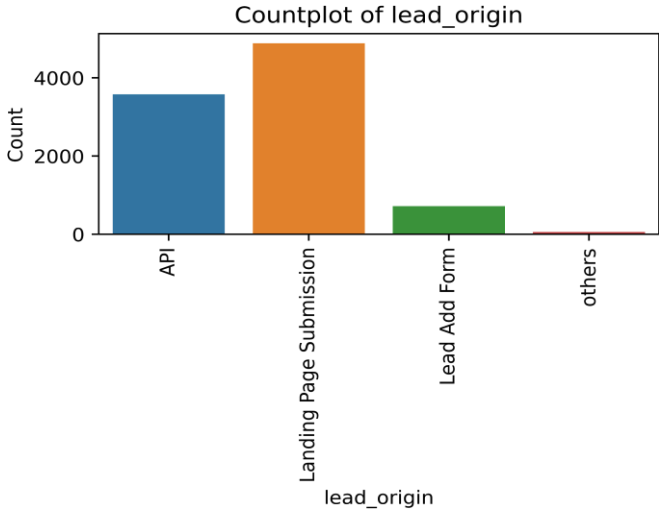


3. PAGE VIEWS PER VISIT: The distribution is right skewed with 0.32 skewness. As per the distribution, maximum leads only visit the page 1-5 times and do not engage with X education. The leads whose number of page visits are the highest are the most engaged customers.

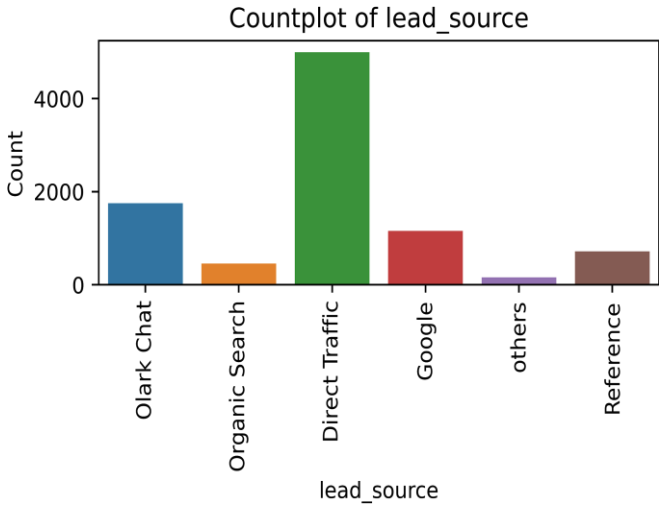


Categorical Columns:

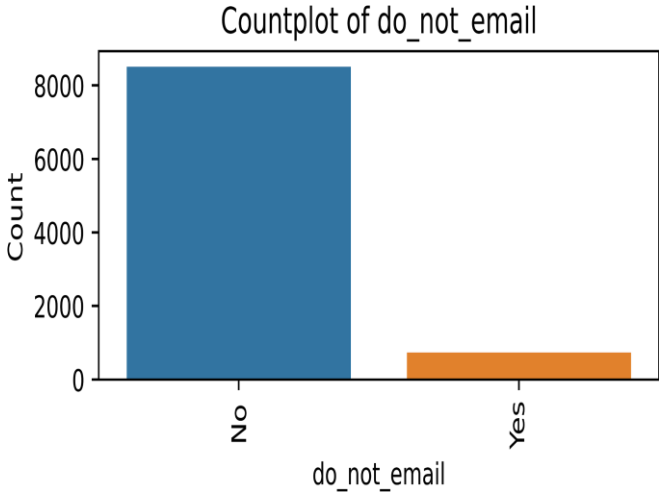
1. LEAD ORIGIN: The maximum number of leads are generated through Landing Page Submission. 'Others' generates the lowest number of leads. The main lead generators are Landing Page Submission (4886) and API (3580).



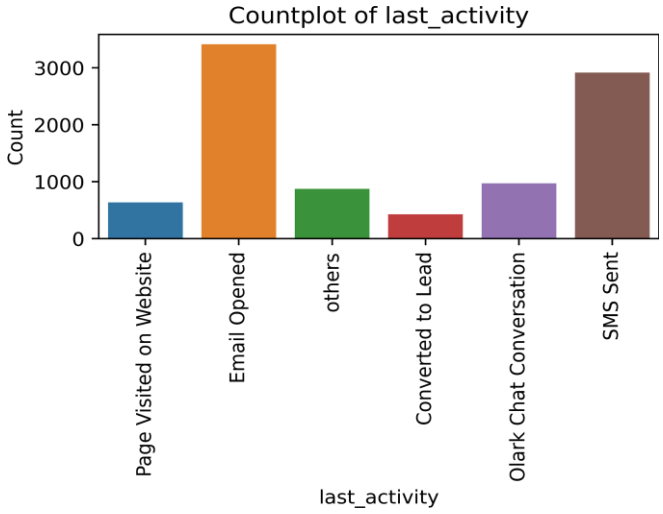
2. LEAD SOURCE: The leads are generated from 6 different sources, highest of which are generated through Direct Traffic (5000) , Olark Chart(1755), and Google(1500).



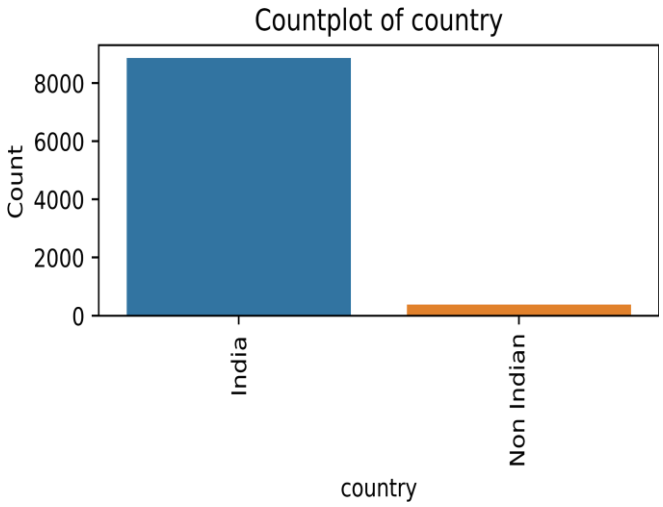
3. DO NOT EMAIL: Majority of the people are ok with receiving email (~92%),people who are ok with email has conversion rate of 40% and people who have opted out of receiving email has lower rate of conversion (only 15%)



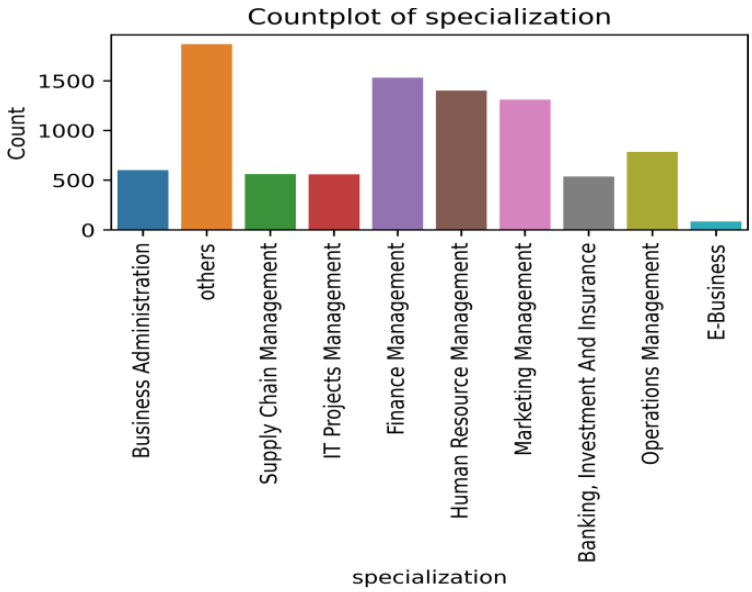
4. LAST ACTIVITY: The last activity by leads was majorly on Email Opened (3437), SMS Sent (2745) and Olark Chat Conversation (973)



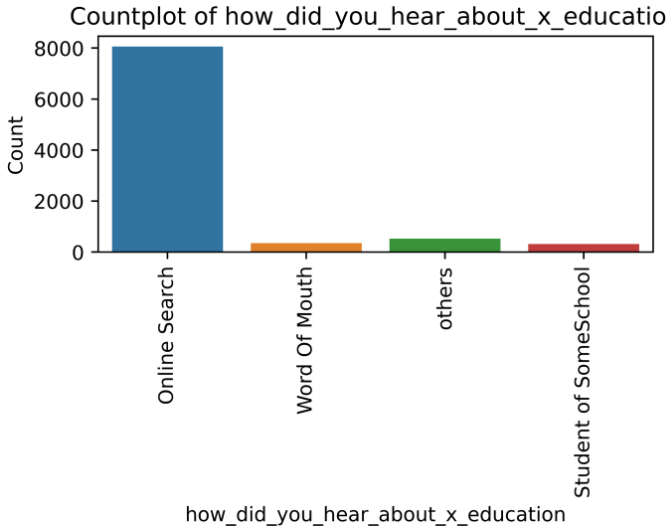
5. COUNTRY: The highest leads are generated from India (8861) and less leads from other countries (387)



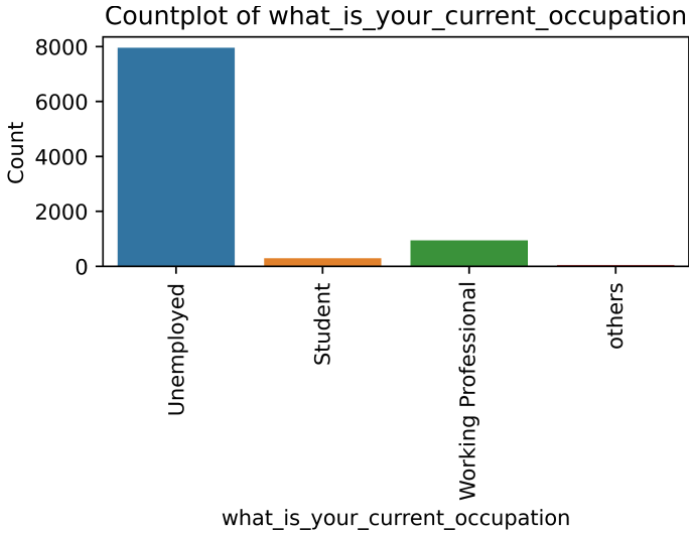
6. SPECIALIZATION: Others and Finance Management is the most common specialization where people have worked before



7. HOW DID YOU HEAR ABOUT X EDUCATION: The maximum leads are generated through Online Search (8058) and Others are by Word of Mouth, Student of Some School and other medium

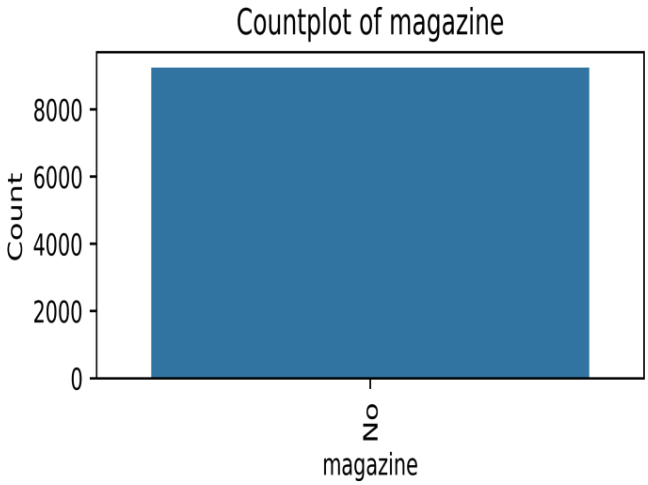


8. WHAT IS YOUR CURRENT OCCUPATION: Maximum leads are generated from people who are unemployed (~85%)

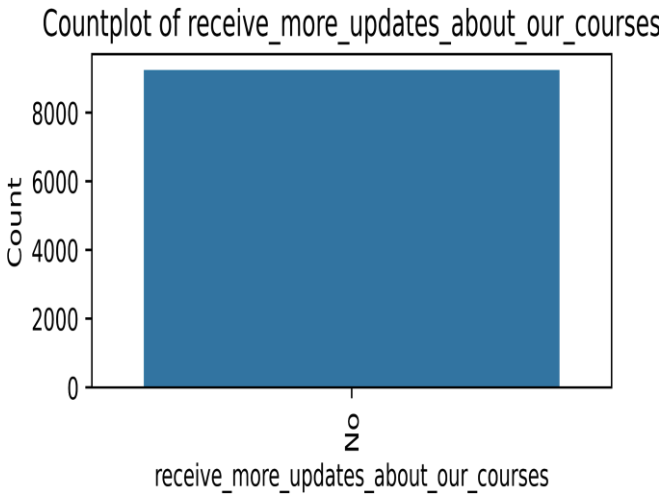


9. **WHAT MATTERS MOST TO YOU IN CHOOSING THIS COURSE:** Maximum leads have mentioned that better career prospect matters the most.

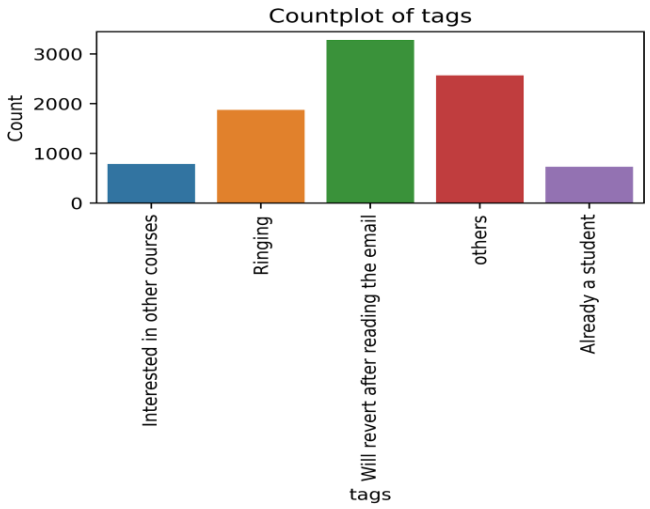
10. **MAGAZINE:** It only contains 'No', so will not generate any insight. Hence, we will drop this column.



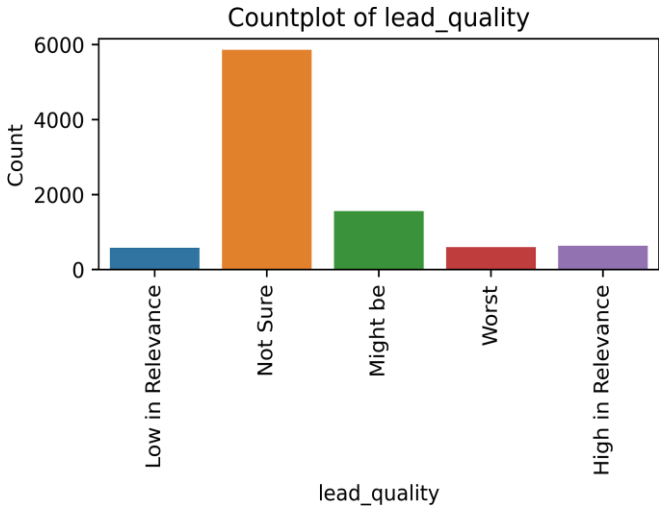
11. **Receive More Updates About Our Courses:** It only contains ‘No’. So will not generate any insight. Hence, we will drop this column.



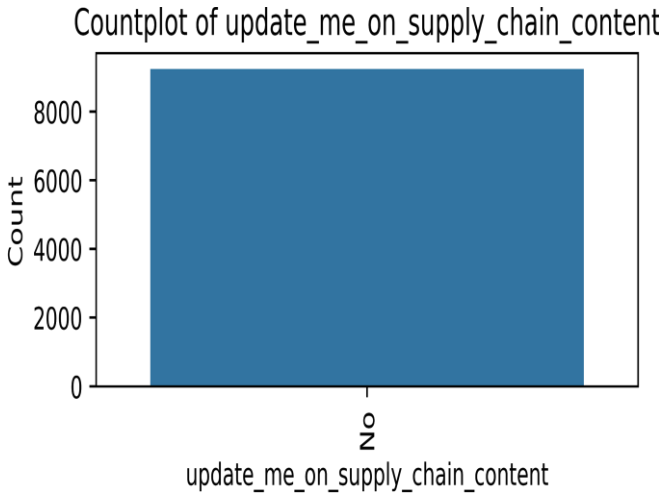
12. **TAGS:** The maximum leads generated do revert after reading email (3170) and ‘Others’.



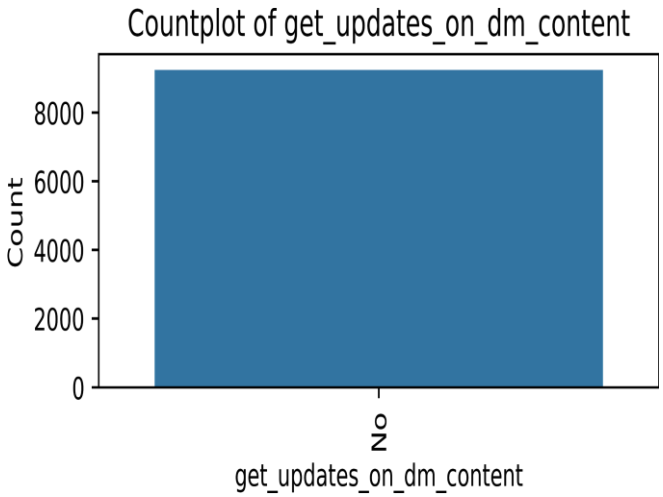
13. LEAD QUALITY: Most of the candidates are assessed as 'Not Sure' (5872), 'Might be' (1760), and
'High in Relevance' (837) leads.



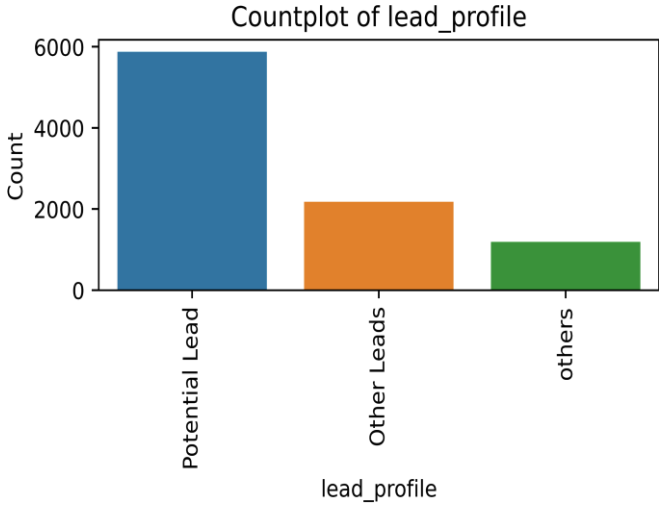
14. UPDATE ME ON SUPPLY CHAIN CONTENT: No customer wants updates on Supply Chain Content. It only contains 'No'. So will not generate any insight. Hence, we will drop this column.



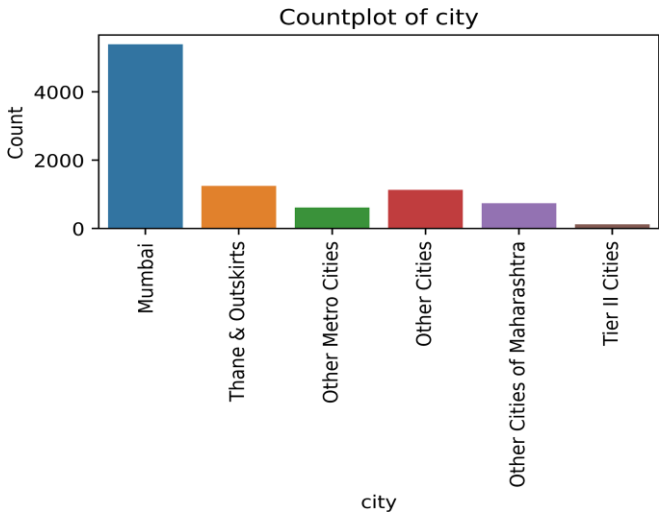
15. GET UPDATES ON DM CONTENT: No customer wants updates on the DM Content. It only contains no. So will not generate any insight. Hence, we will drop this column.



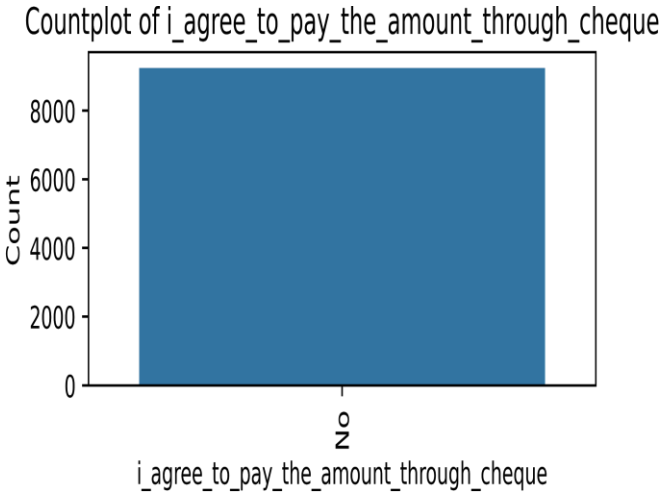
16. LEAD PROFILE: A lead level assigned to each customer based on their profile. 5875 lead is highest in number and have been assigned as 'Potential Lead'



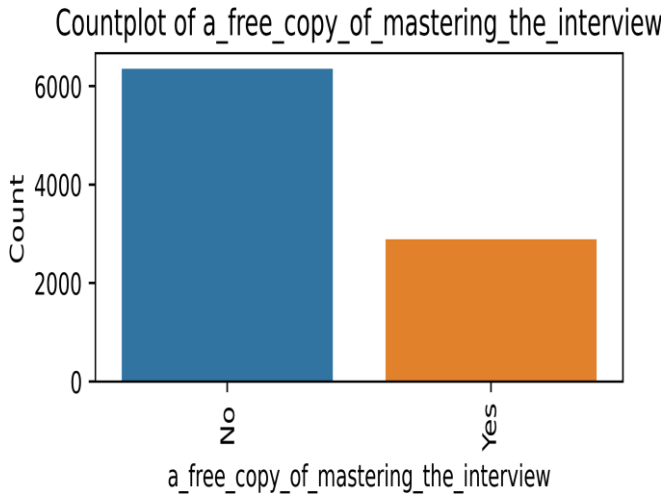
17. **CITY:** Most of the customers belong to Mumbai (6891)



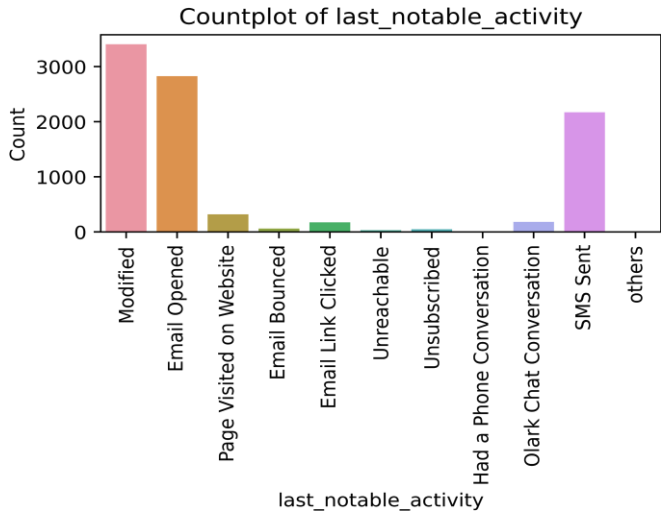
18. **I AGREE TO PAY THE AMOUNT THROUGH CHEQUE:** Nobody has agreed to pay the amount through cheque. Hence, we will drop this column



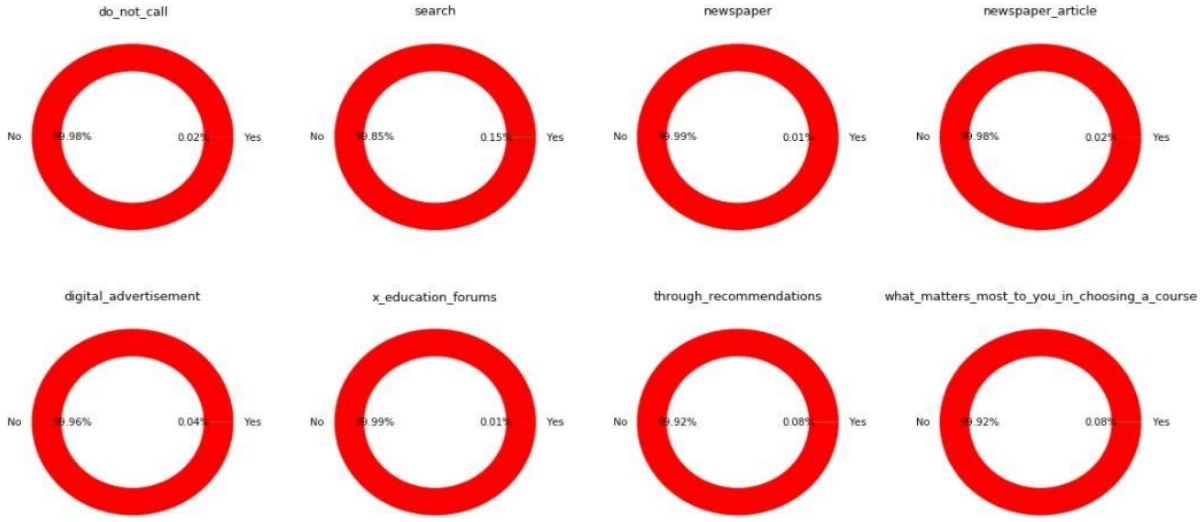
19. A FREE COPY OF MASTERING THE INTERVIEW: Most leads (~68%) opt out of free copy of the book



20. **LAST NOTABLE ACTIVITY:** Most notable activities of the leads were account 'Modified'(36.8%), 'Email Opened'(30.5%) and 'SMS Sent' (23.5%).



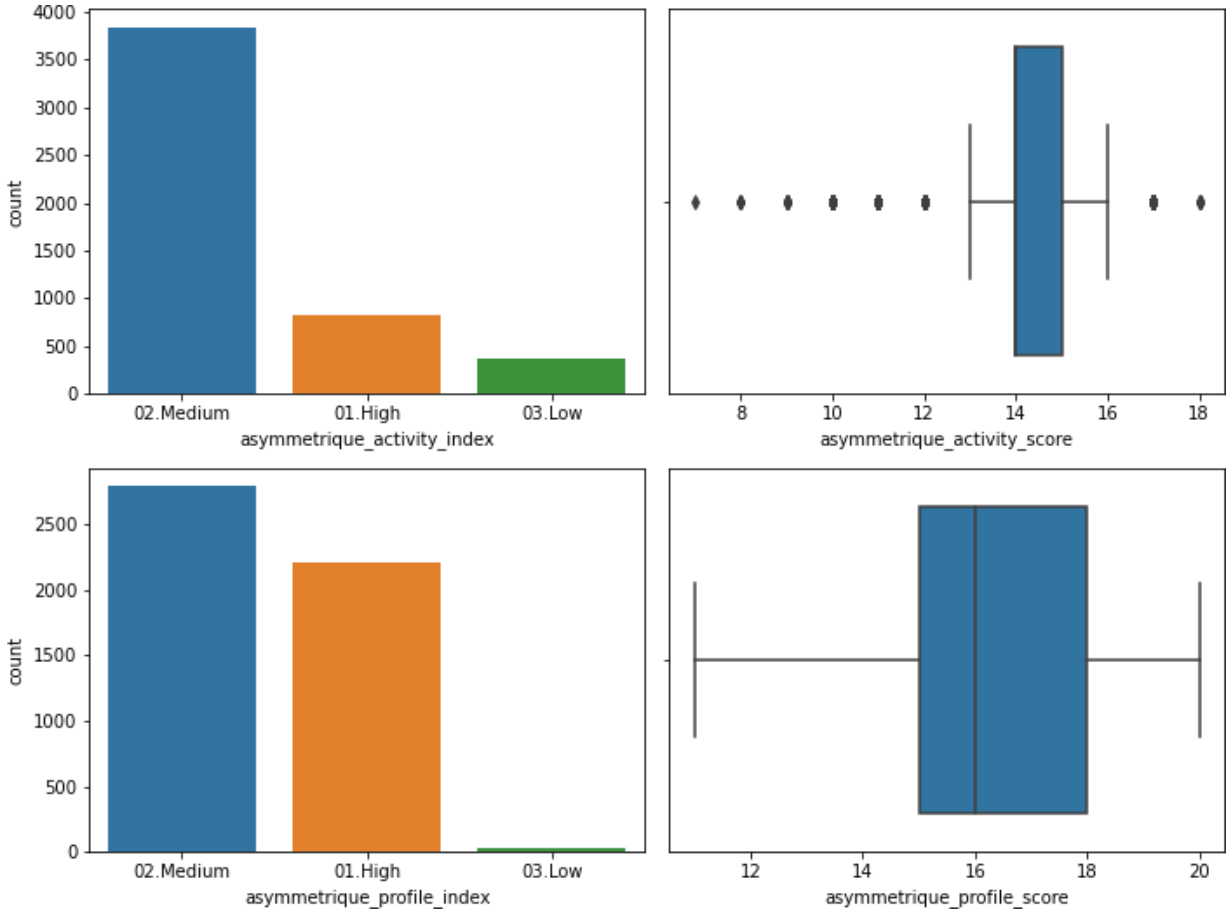
21. Do Not Call, Search, Newspaper, Newspaper Article, Digital Advertisement, X Education Forums, Through Recommendations: Maximum values are the same, so we cannot generate any insight for lead conversion from this data. Hence, we will drop these columns.



Asymmetrique Scores and Index:

Asymmetrique Activity Index and Asymmetrique Activity Score: They are assigned once a lead is assessed by X education on basis of other features like Total Visits, Total Time Spent on Website, Page Views Per Visit, Last Activity

Asymmetrique Profile Index and Asymmetrique Profile Score: They are assigned once a lead is assessed by X education on basis of other features like specialization, current occupation, and what matters most in choosing a course



Inferences:

1. Leads with 01. High and 03. Low activity index is low in numbers as these leads generally do not spend too much time on the website. Maximum leads spend a good amount of time on the website and are of index 02. Medium. The combination of activity actions is assigned a numerical value for each lead. As the index 02. Medium has the highest count, consequently, activity scores lie in 13 to 16. The 01. High index leads generate the outliers above the IQR and 03. Low index leads generate the outliers below the IQR

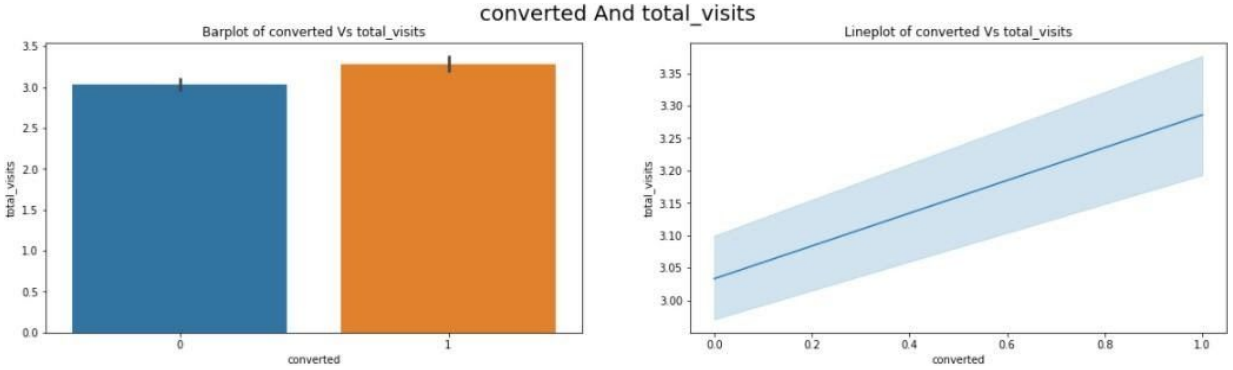
2. Leads with 03. Low profile index is low in numbers as we have few leads coming from a non- relevant background. Generally, people with relevant background visit the website to upskill and are indexed as 01. High and people with semi-relevant background visit the website to garner more skills. The combination of profile features of a lead is assigned a numerical value. As the index 02. Medium has the highest count, consequently, activity scores lie in 15 to 18. 01.High index profiles are having a score above 18

NOTE: 45% of the records do not have these values assigned, so we need to drop these columns as it is not reliable to impute any value in it or do further exploration of this data.

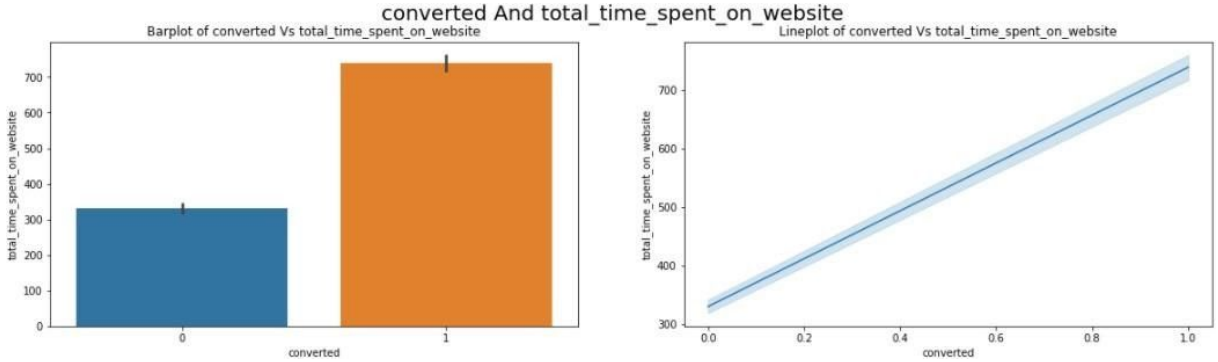
Bivariate Analysis:

Numerical vs Target:

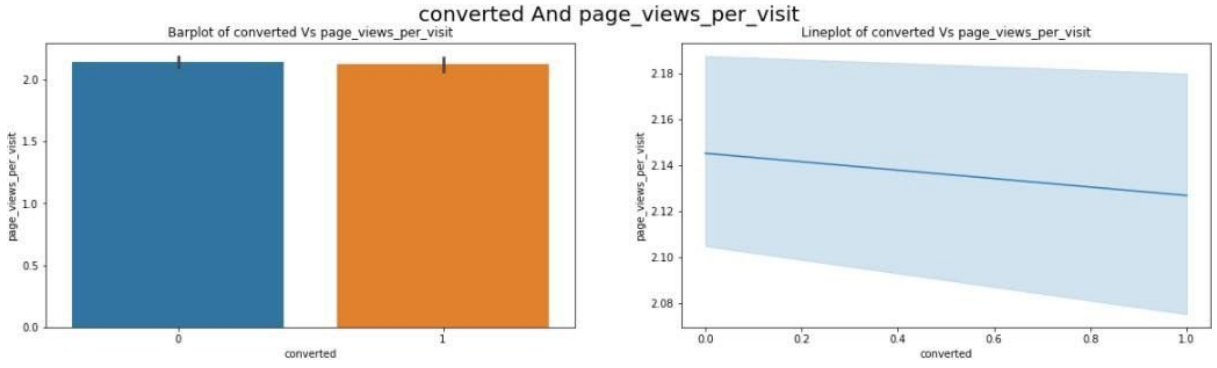
1. **converted vs total_visits**: The probability of a lead being converted does not really depend on the total number of visits made by the customer on the website.



2. **converted vs Total_time_spent_on_website**: The more the amount of time a person spends on the website, the probability of that person being converted to a customer is more.

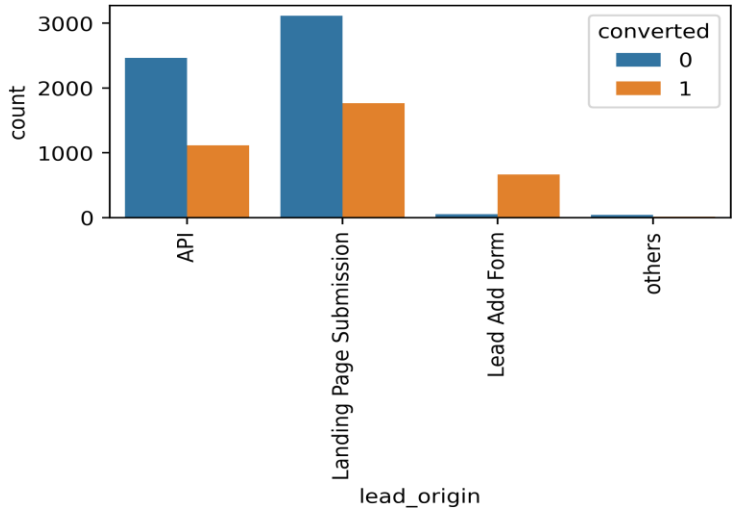


3. **converted vs Page_views_per_visit**: The probability of a lead being converted into a customer increases with a small decrease of Page_views_per_visit from its mean value.

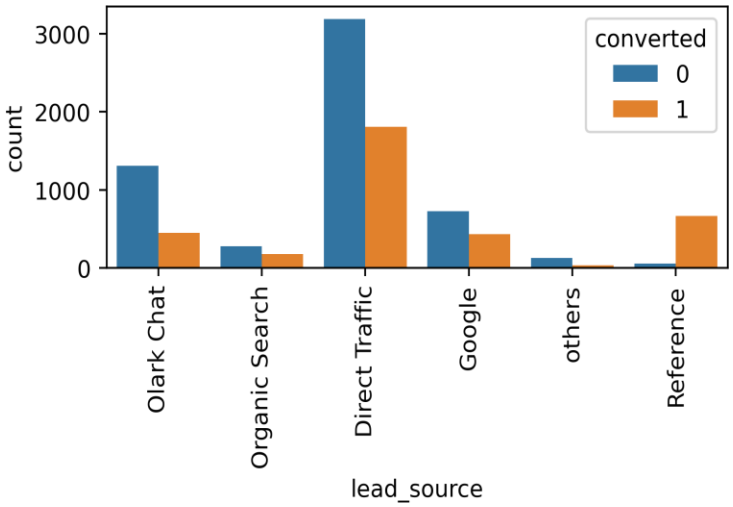


Categorical vs Target:

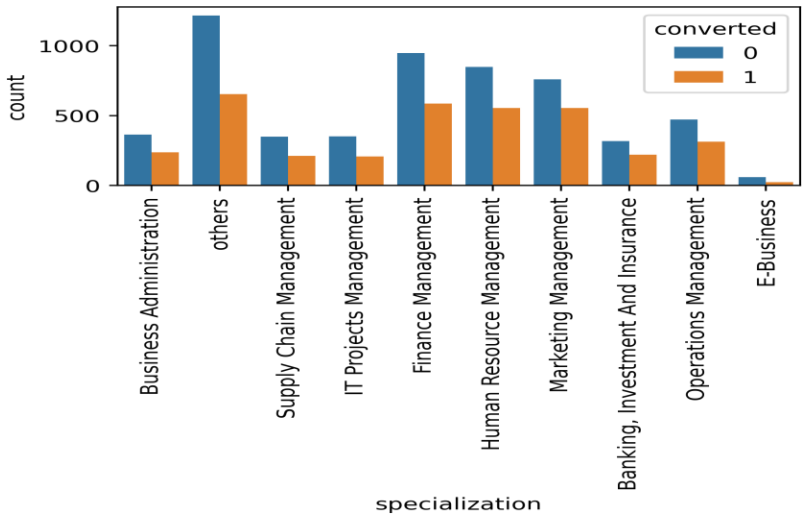
1. **Lead Origin**: API and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable. Lead Add Form has more than 90% conversion rate but count of lead are not extremely high. Others are very less in count.



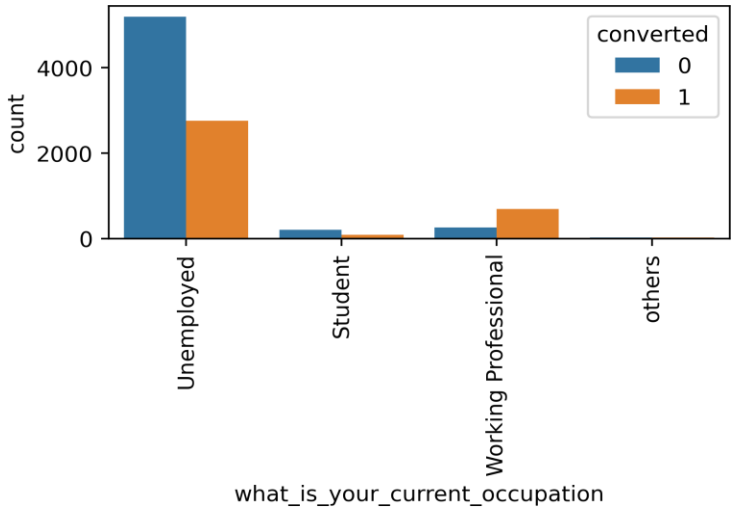
2. **Lead Source**: Olark Chat and Direct traffic generates maximum number of leads. Conversion Rate of Reference leads is high.



3. **Specialization**: Focus should be more on the Specialization with high conversion rate Finance Management, Human Resource Management and Marketing management has greater than 50% conversion rate. While for others count is high, we have a conversion rate of less than 35%



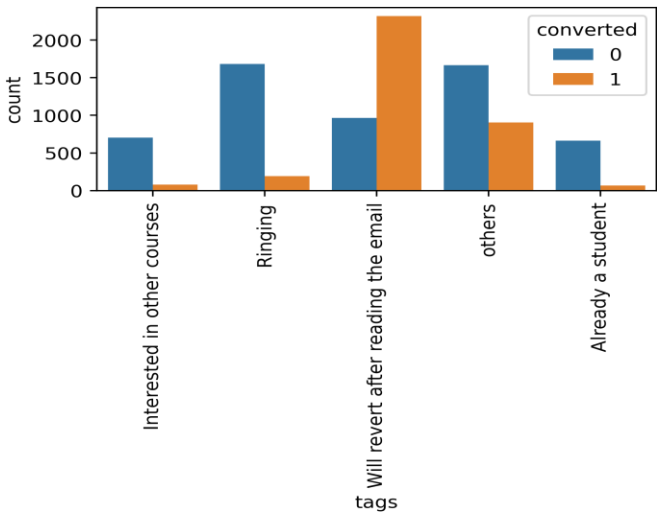
4. **What is your current occupation**: 80% conversion rate of Working Professional. Though Unemployed people have been contacted in the highest number, the conversion rate is low
~40%.



5. **How did you hear about X Education**: “Student of SomeSchool”, “Word Of Mouth” and “Others” has the highest conversion rate of 45%. While for Online Search count is high, we have a conversion rate of less than 35%

6. **Last Activity**: Most of the leads have their Email opened as their last activity. Conversion rate for leads with last activity as SMS Sent is almost 60%. Conversion rate for leads with last activity as Email Opened is almost 50%.

7. **Tags**: Will revert after reading email has ~60% conversion rate. ‘Others’ has ~40% conversion rate. Ringing has extremely low conversion rate as leads have not responded on calls at first touchpoint call

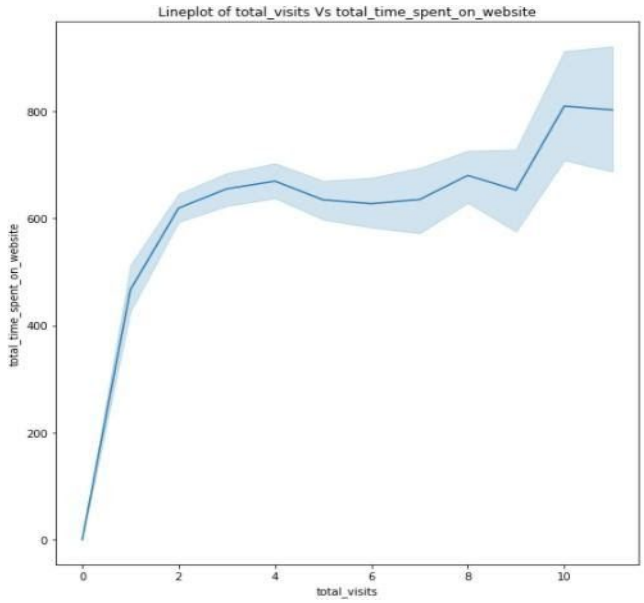


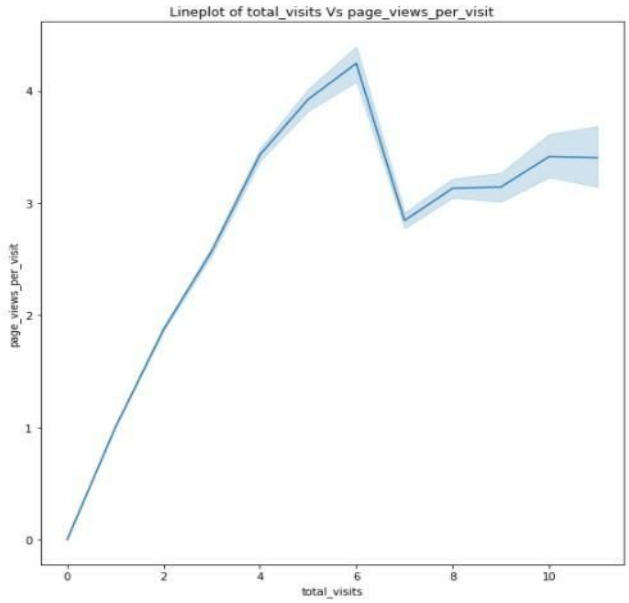
8. **Lead Profile**: Potential Leads have highest conversion of 50%. Other Leads and Others have ~40% conversion rate.

9. **Last Notable Activity**: SMS Sent has the highest conversion rate of ~70%. Email opened has a good conversion rate of ~40%. While for Modified count is high, we have a conversion rate of ~20% as leads are modifying their accounts and not coming back to the website.

Numerical vs Numerical:

1. The total number of visits made by the customer on the website lies between 0 to 11 the Average number of pages on the website viewed during the visits is anywhere in between 0 - 3.5
2. Those customers who tend to spend anywhere between 0 - 500 minutes on the website visit around 0 to 11 number of pages on the website.

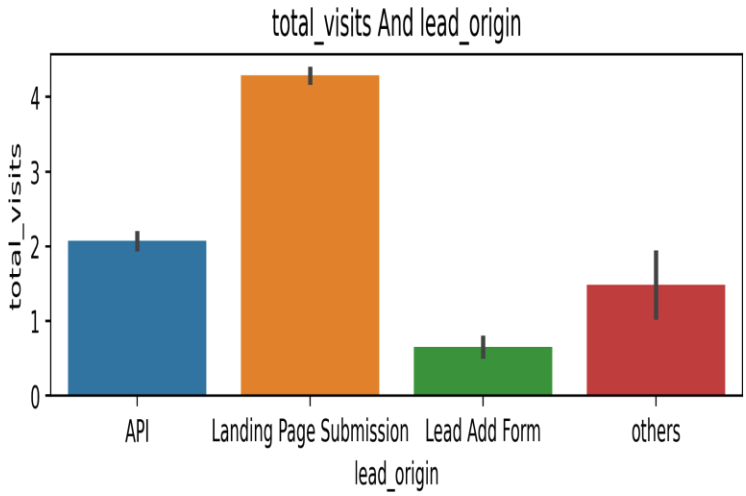




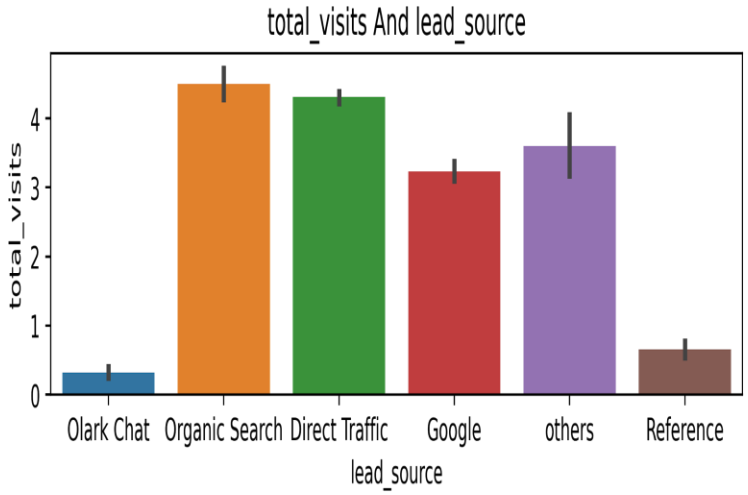
Numerical vs Categorical:

1. Total Visit vs categorical variables:

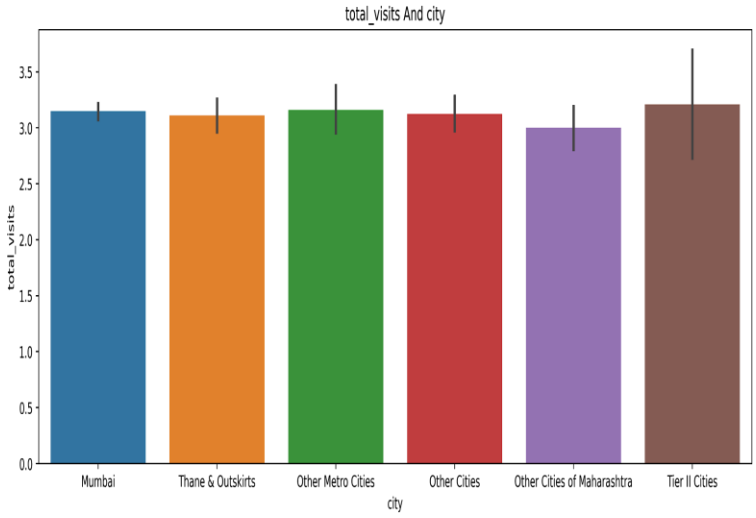
a. *Lead Origin:* API, Landing Page submission, and Others contribute as the highest source of Total visits on the website.



b. *Lead Source:* Customers through Organic Search and Direct Traffic have visited the websites most.



c. City: Customers from different cities have visited the website around 3 times, the highest being from Tier II Cities.



d. *How did you hear about X Education:* Students who spent a significant amount of time on the website heard about X Education by Word Of Mouth, Student of Some School, and others.

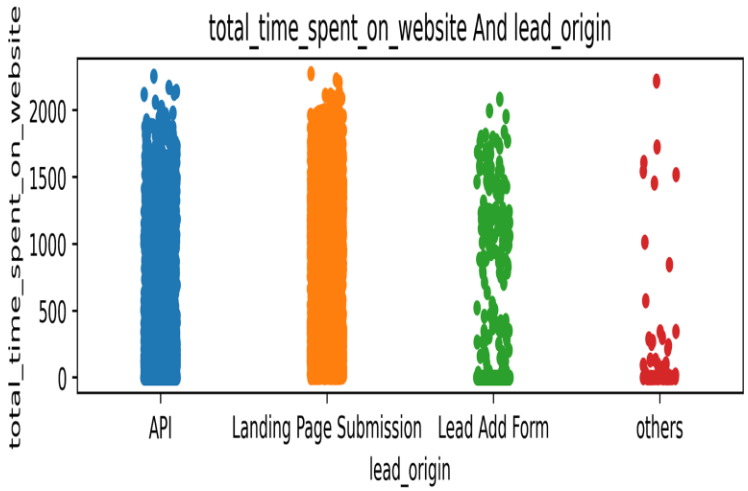
e. *What is your current occupation:* Most of the leads who have visited the website are unemployed (5600) or Others which consists of other occupations in small counts.

f. *Tags:* People who spent a significant amount of time on the website Will revert after reading the email or not have responded on touchpoint call.

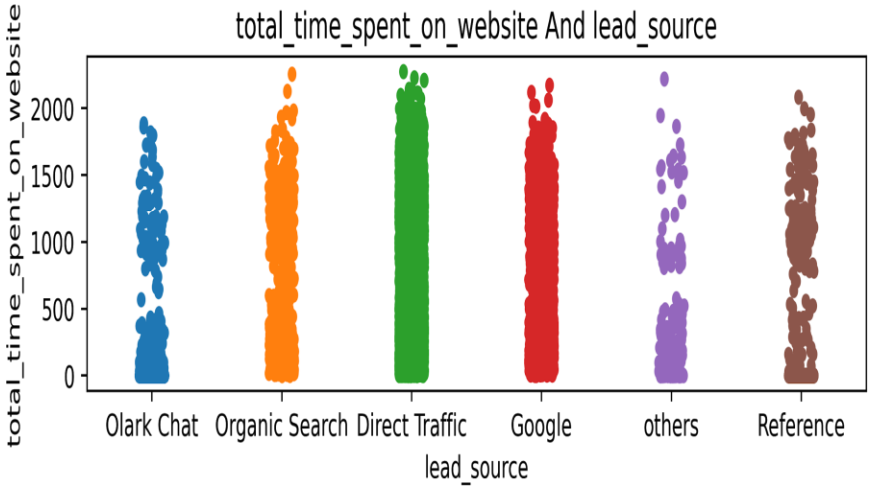
g. *Lead Quality:* People who have visited the max number of times are assessed as relevant leads.
h. *Lead Profile:* People who have visited the X education website the maximum number of times on the website are “Potential leads” and “other leads”.

2. Total Time Spent on Website vs categorical variables:

a. *Lead Origin:* Leads from Landing Page Submission and API spend the highest time on the website.



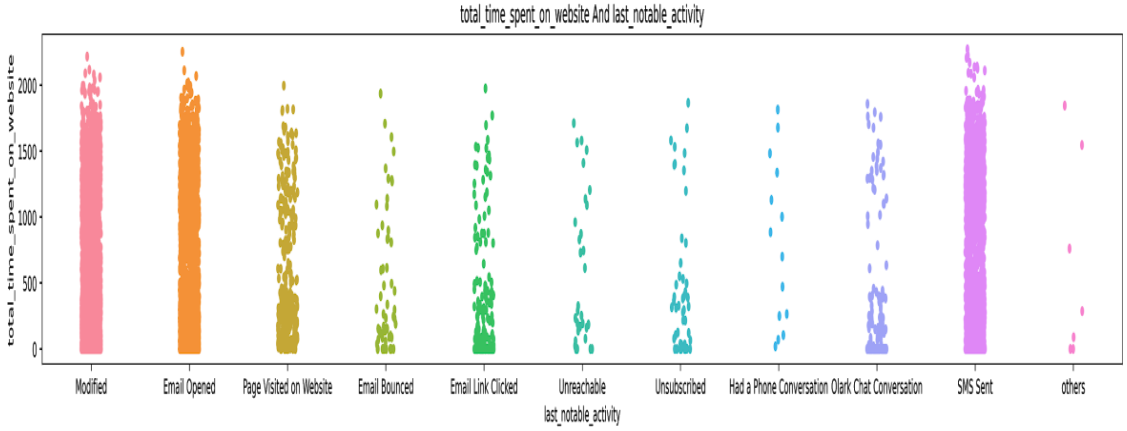
b. *Lead Score:* Organic Search, Direct Traffic and Google students spend the highest time on the website.



c. *last Activity:* In reference to the time spent on the website the student's most specific activity was SMS Sent, Page Visited on Website, and Email Opened.

d. *City:* Leads from Mumbai have spent high time on the website.

e. *Last Notable Activity:* Leads who have spent high time on the website notable open Emails, send SMS, and Modify accounts.



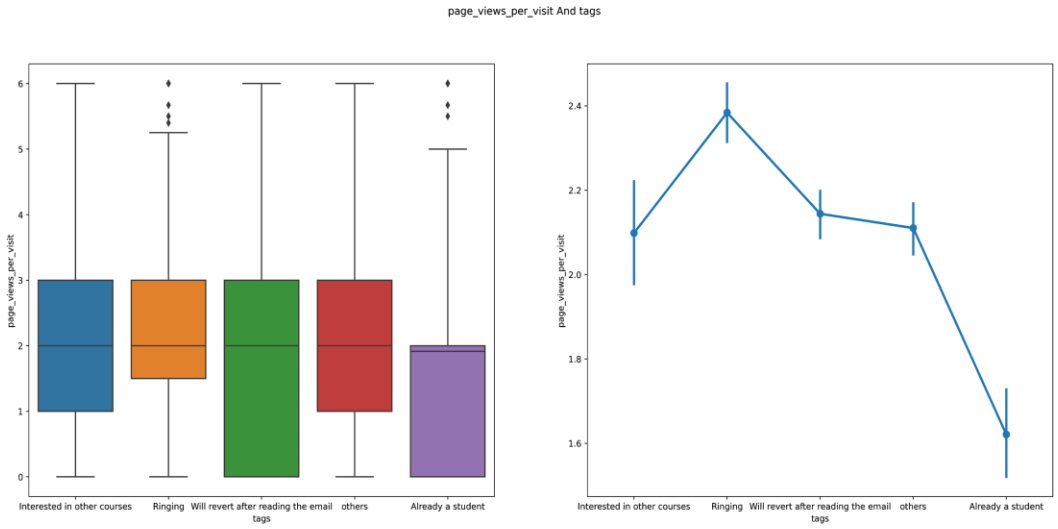
3. Page views per visit vs categorical variables:

a. *Lead Origin:* Leads from Landing Page Submission spend the highest time on the website.

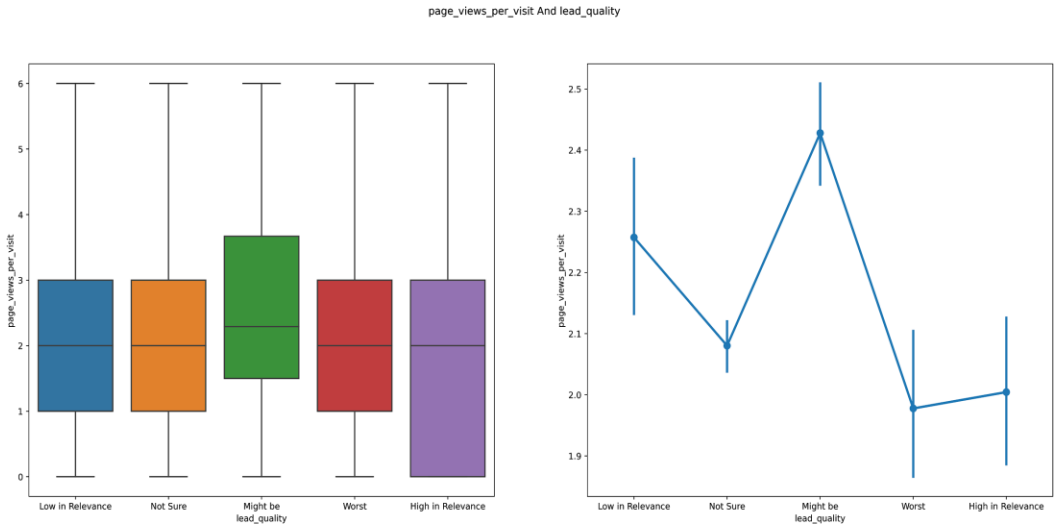
b. *Lead Score:* Direct Traffic generates the highest views

c. *What is your current occupation:* Unemployed and Working Professional generate lower views. Other occupation has generated the highest views

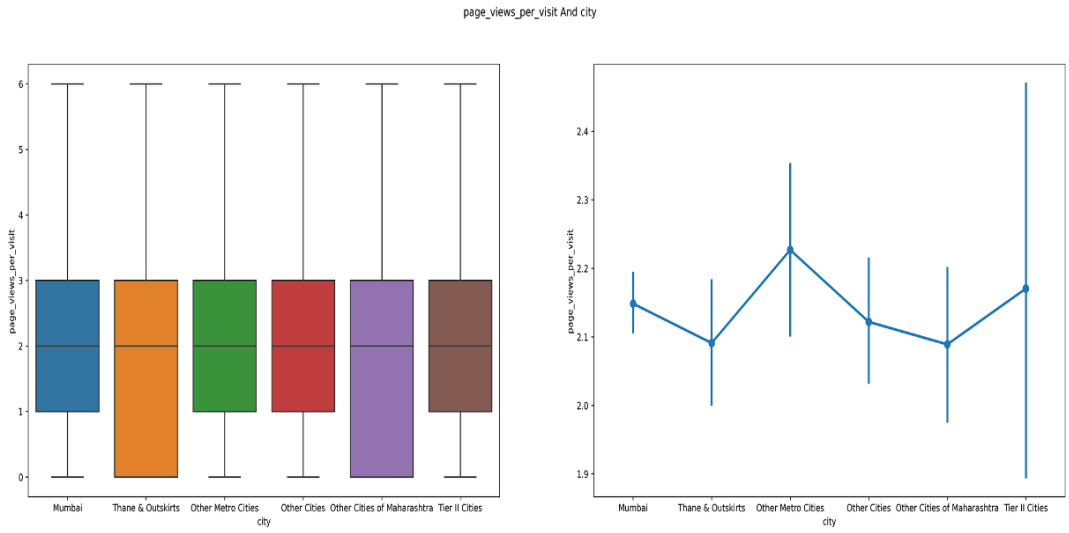
d. *Tags:* Will revert after reading mail generates lower views. Touchpoint call/Ringing has generated the highest views



e. *Lead Quality:* Leads who generate high page visits, are generally who compare between courses, are assessed as 'might be'.

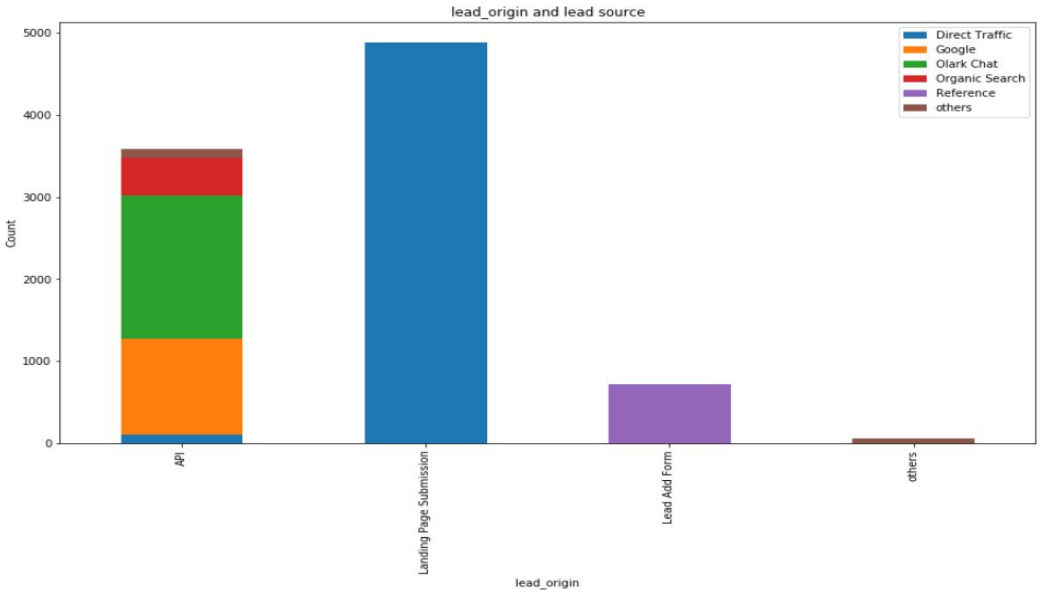


f. *City:* Leads from Mumbai generate lower views. 'Other cities' has generated the highest views

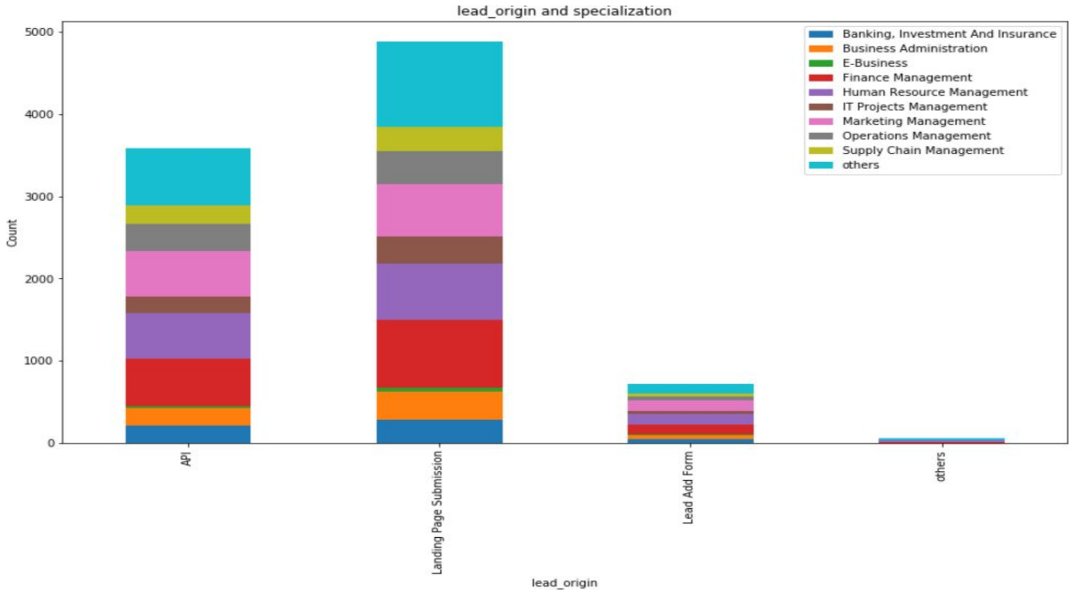


Categorical vs Categorical:

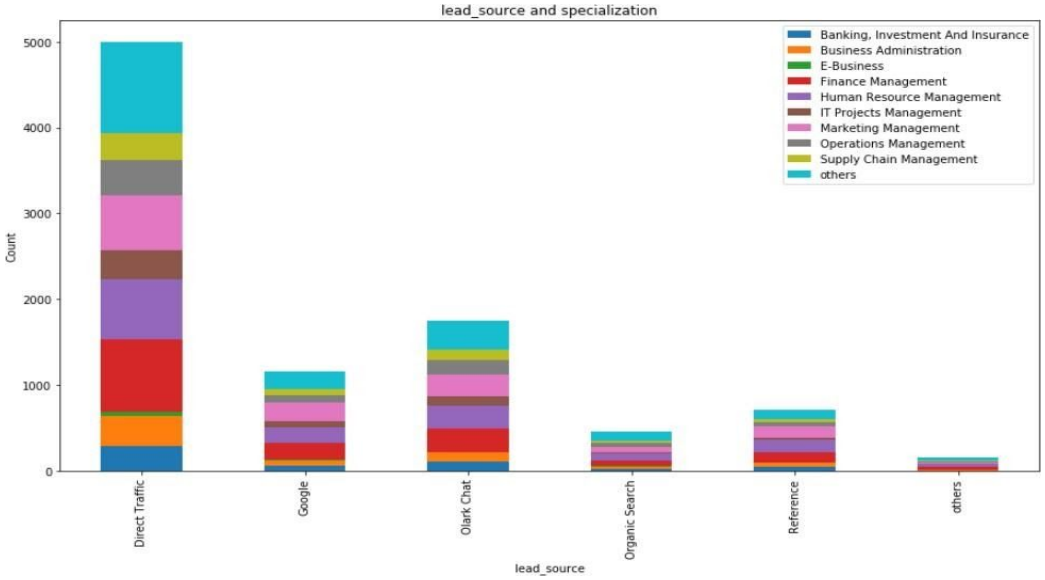
1. Lead_Origin vs Lead_Source: Majority of the leads were from Landing page submissions and API with the source for leads being Google, Direct Traffic to X education website and from sources like Olark chat and organic search



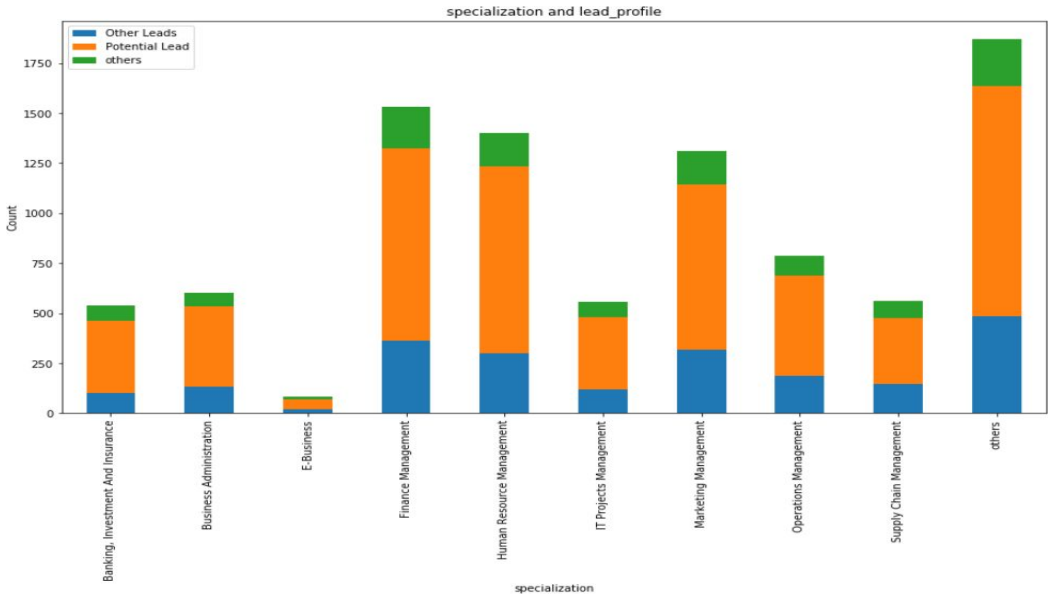
2. Lead_Origin vs Specialization: Majority of the leads who were from Landing page submissions and API specialize in Human Resource Management and Marketing Management and highest in Others



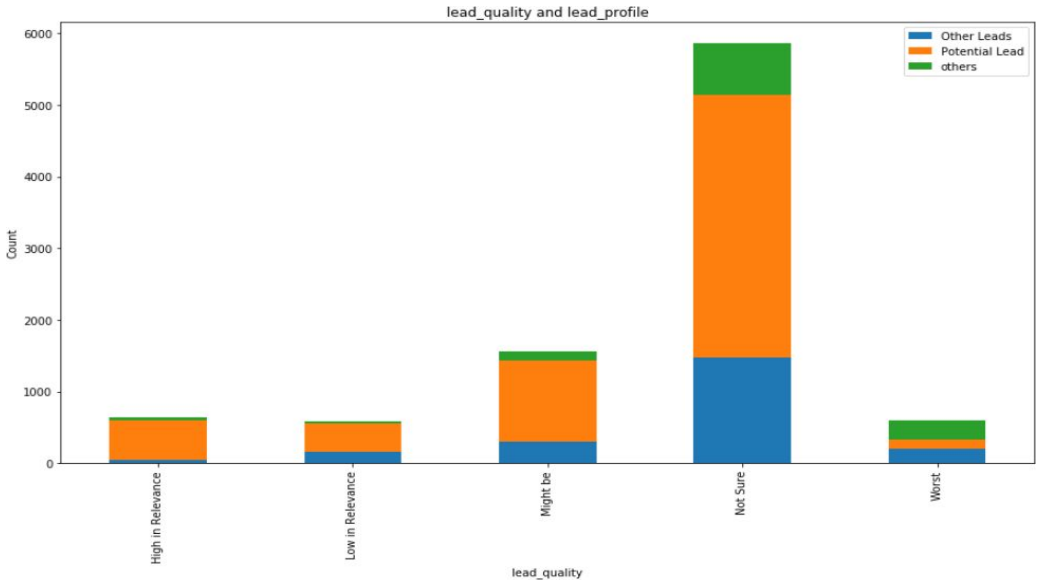
3. Lead_Source vs Specialization: Majority of the lead source were from Direct Traffic, Olark Chat and Google specialize in Finance Management and Human Resource Management but highest numbers is for Others



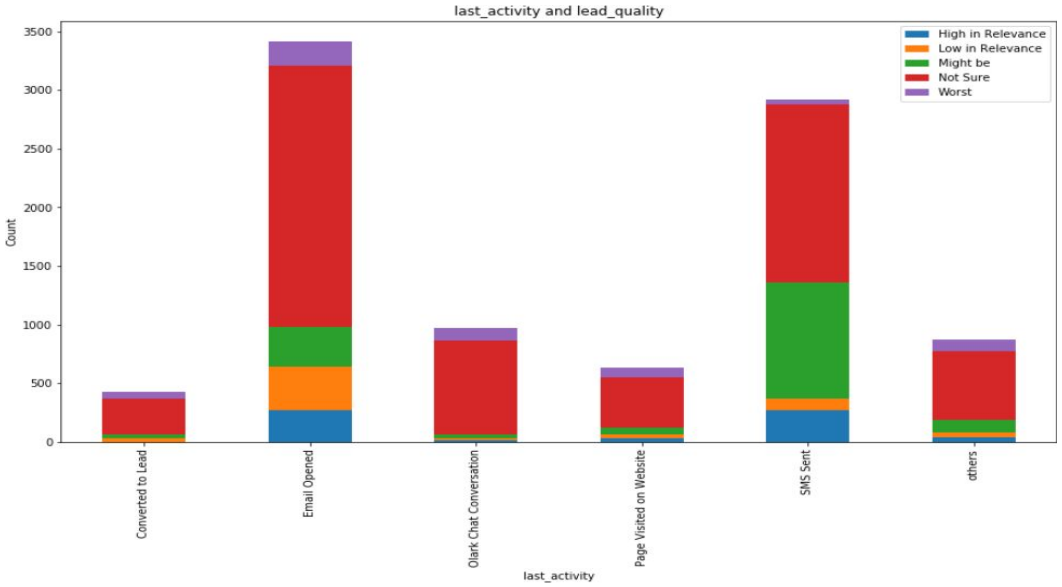
4. Specialization vs Lead_Profile: Majority of the lead have previously worked before in Finance Management, Human Resource Management and Marketing Management, but highest number of leads have worked in Others domain and are assigned as Potential Lead and few as Other leads



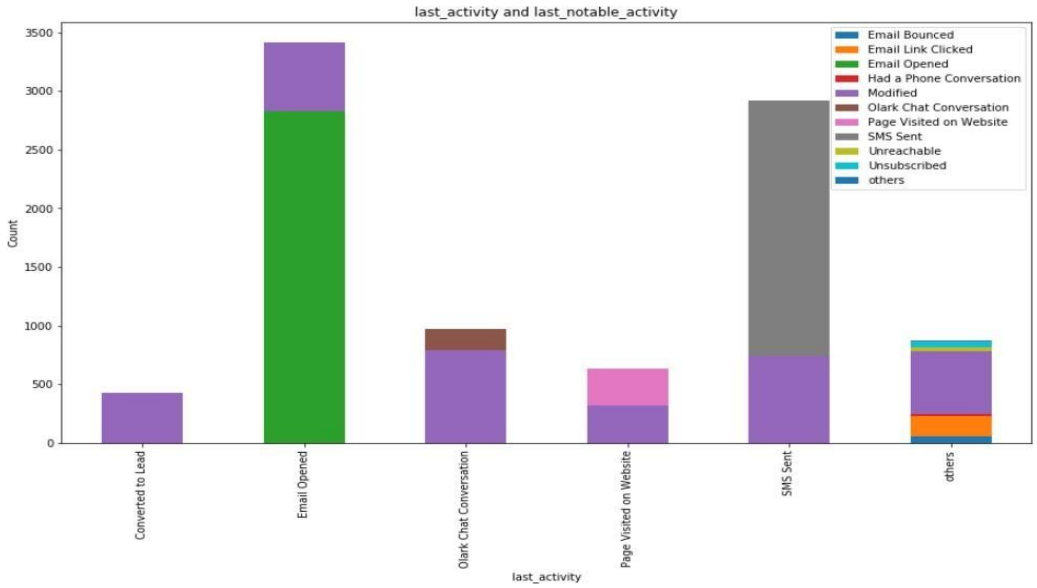
5. Lead_Quality vs Lead_Profile: Maximum leads are assessed as Not sure but assigned as Potential Lead and few as other leads



6. Last_Activity vs Lead_Quality: Most leads who opened their Email were assigned lead quality as Not Sure and who Sent SMS were assigned Not Sure and Might Be

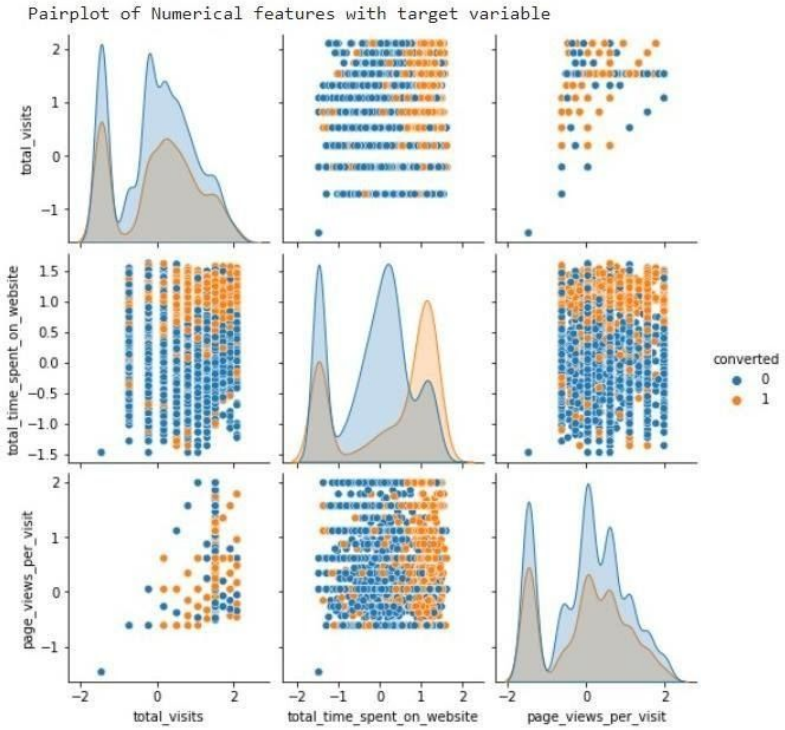


7. Last_Activity vs Last_Notable_Activity: Leads generally Modified account after Email opening, sending SMS or Converted to Lead



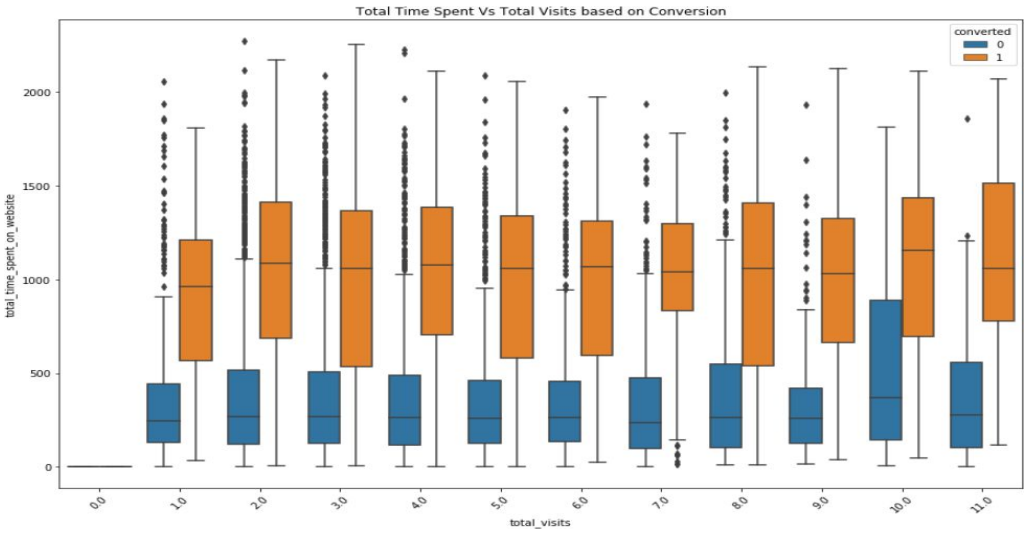
Multivariate Analysis:

1. **Pair-plot for multivariate analysis**:

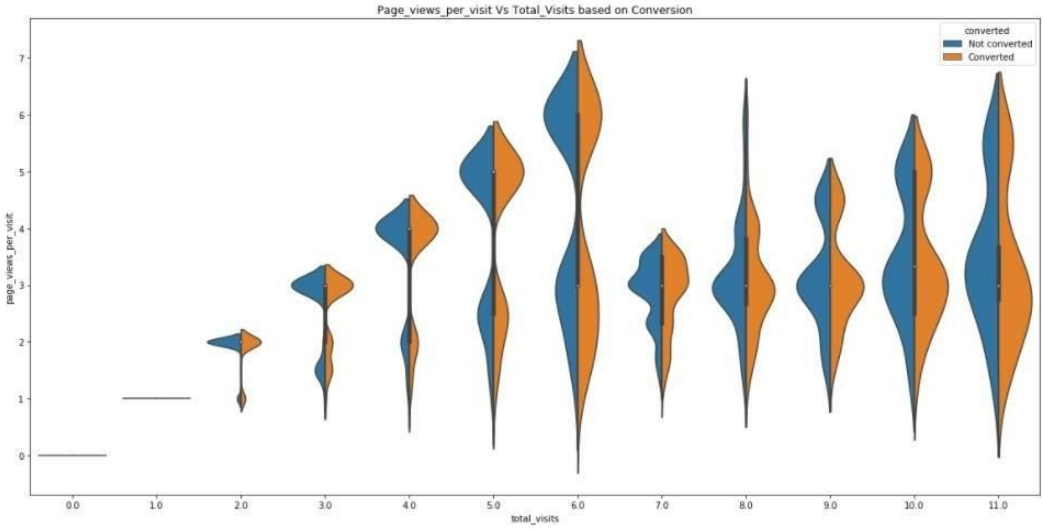


- There is no separation of converted and unconverted leads with respect to total_visits and Page_views_per_visit
- We see a significant separation of converted and unconverted leads with respect to Total_time_spent_on_website

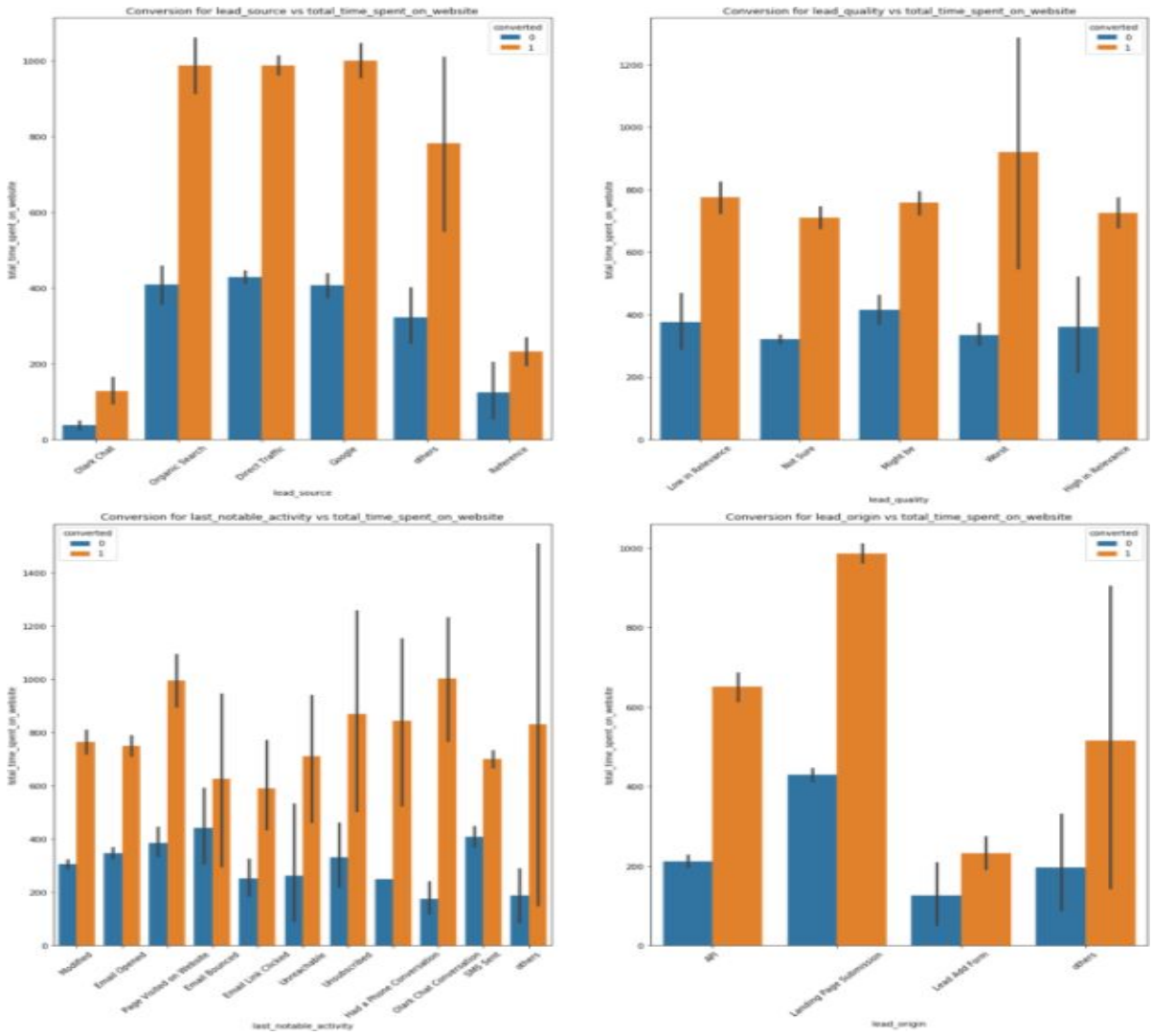
2. **Numerical variables with target**:



- The highest number of conversions happen when people are spending around 18 hours or above on the website.
- People who spent around 3 hours on the website didn't opt for any courses. From the boxplot we can see better that the longer you stay on the website, the higher your chances of conversion as well.

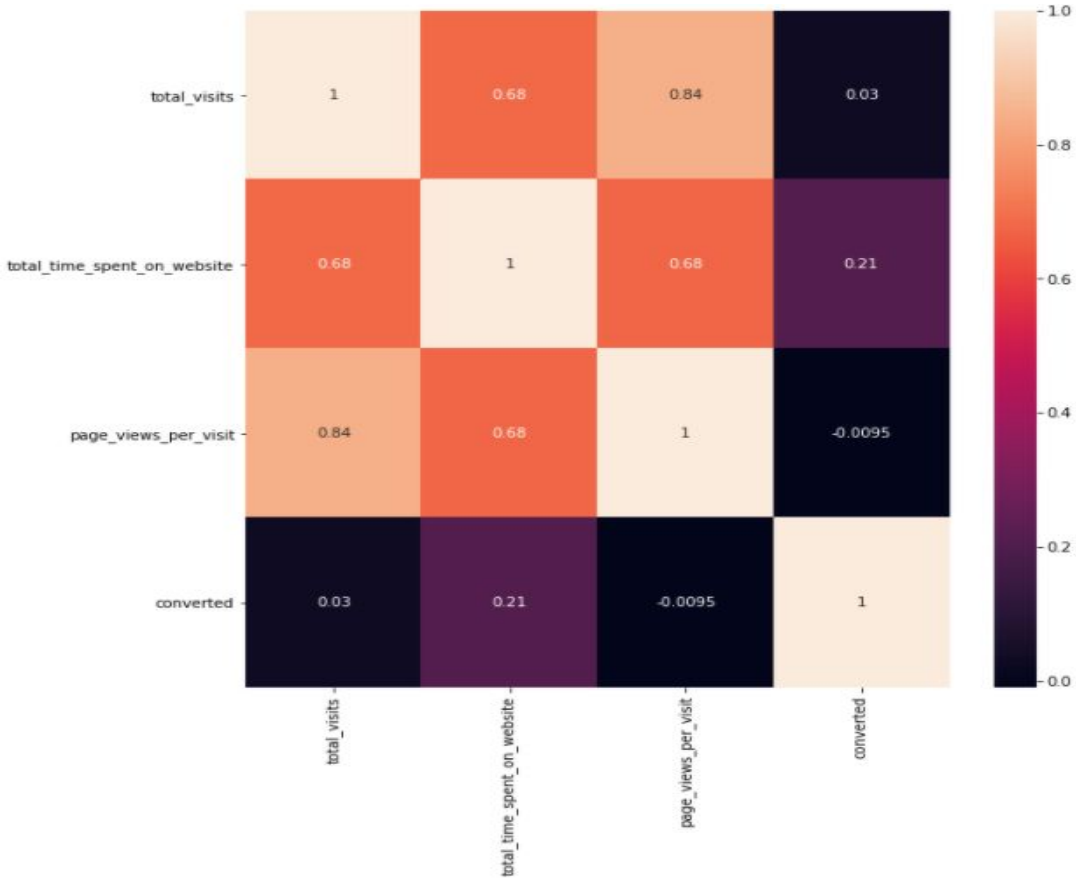


- Most total visits on the website is 6. Until total visits is 6, the conversion rate is higher than when page views per visit is high. When total visits cross 6, the conversion rate drops as total pageviews decrease



- The rate of increase in 'conversion rate' increases with more time spent on the website irrespective of Lead Quality assessment, leads get converted in equal ratio.
- Leads who opened website pages as well as spent time on the website got converted more.
- Landing Page submission generated the maximum engagement and consequent conversion

3. **Heatmap for multicollinearity**:



- There is moderate multicollinearity between total_visits and Total_time_spent_on_website of 0.68 strength
- There is high multicollinearity between total_visits and Page_views_per_visit of 0.84 strength

Statistical Significance of Features

T-Test for numerical variables:

We have performed T-Test for independence testing of the numerical variables, with a 95% confidence interval. Therefore, with a 5% significance level (at p-value 0.05), we reject the null hypothesis because the p-value is greater than the critical value.
NULL HYPOTHESIS(H0): The feature is statistically significant.
ALTERNATE HYPOTHESIS(H1): The feature is statistically insignificant.

Feature name	p-value	Significance
Total_visits	9.06e-6	Yes
Total_time_spent_on_website	6.06e-285	Yes
Page_views_per_visit	0.59	No

Chi Square test for categorical variables:

We have performed Chi Square test of independence for the categorical variables, with a 95% confidence interval. Therefore, with a 5% significance level (at p-value 0.05), we reject the null hypothesis because the p-value is greater than the critical value.
NULL HYPOTHESIS(H0): The feature is statistically significant.
ALTERNATE HYPOTHESIS(H1): The feature is statistically insignificant.

Features	P-Value	Significance
Lead_Origin	3.29e-212	Yes
Lead_Source	6.43e-223	Yes
Do_Not_Email	1.33e-38	Yes
Last_Activity	0	Yes
Country	0.011	Yes
Specialization	0.003	Yes
How_did_you_hear_about_X_Education	0.004	Yes

What_is_your_current_occupation	4.25e-114	Yes
What_matters_most_to_you_in_choosing_a_course	0.87	No
Tags	0	Yes
Lead_Quality	0	Yes
Lead_Profile	1.39e-40	Yes
City	0.18	No
A_free_copy_of_Mastering_The_Interview	0.00014	Yes
Last_Notable_Activity	5.55 e-279	Yes

Anderson-Darling for numerical variables normality distribution:

We have performed Anderson-Darling test of normality for the numerical variables, with a 95% confidence interval. Therefore, with a 5% significance level (at p-value 0.05), we reject the null hypothesis because the p-value is greater than the critical value.

NULL HYPOTHESIS(H0): The data comes from a normal distribution.
ALTERNATE HYPOTHESIS(H1): The data does not come from a normal distribution.

1. Total_visits: The calculated p-value is 0.787. Hence, this feature is not normal
2. Total_time_spent_on_website: The calculated p-value is 0.787. Hence, this feature is not normal
3. Page_views_per_visit: The calculated p-value is 0.787. Hence, this feature is not normal

Feature Engineering

Feature Generation

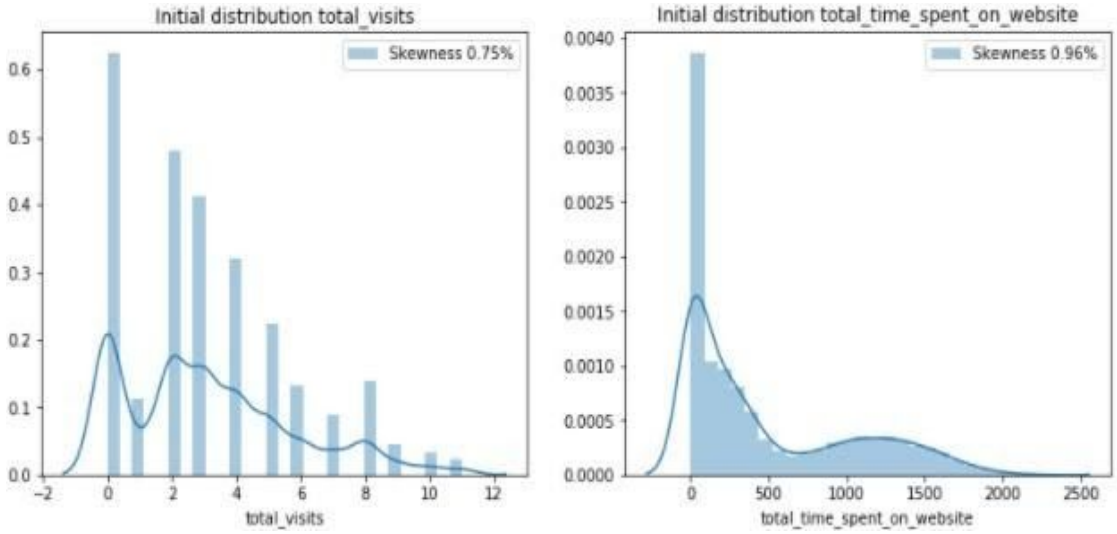
We have generated a new feature ‘time_spent_per_visit’ to reduce the number of features.

```
vif_df_LS2['time_spent_per_visit']=vif_df_LS2['total_time_spent_on_website']/vif_df_LS2['total_visits']
```

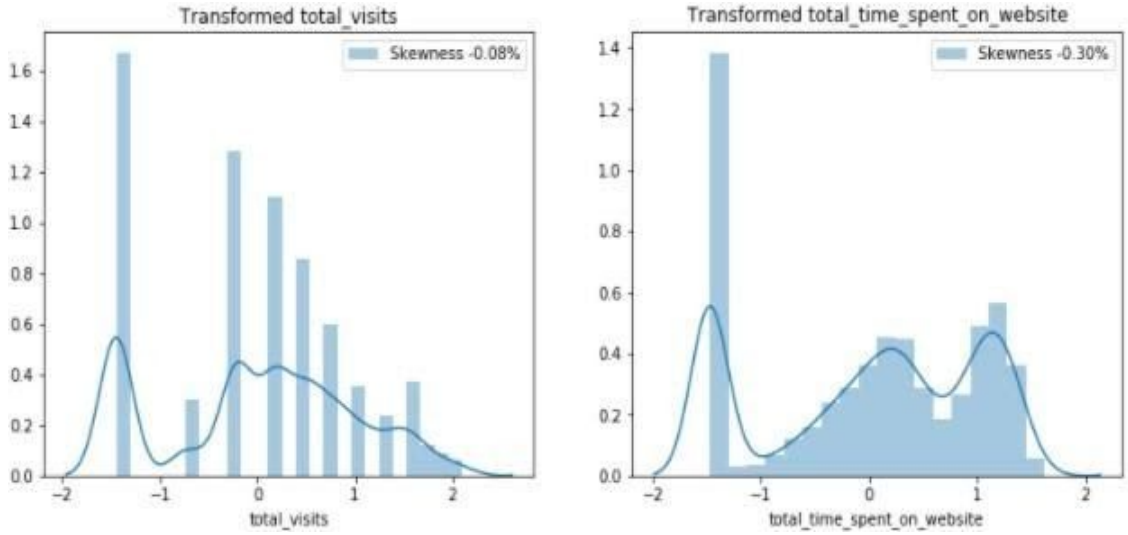
The new feature is significant towards target variable as per the T-Test conducted where p-value of the feature is less than 0.05.

Transformation and Scaling

Before transformation and scaling, the skewness was



We have used Power Transformation for scaling and transformation which generates the best result for this dataset. Transformation and Scaling done on numerical variable and skewness reduced.



Dummy Variable Encoding

One Hot Encoding is a process in the data processing that is applied to categorical data, to convert it into a binary vector representation for use in machine learning algorithms to leverage the information contained in a category value without the confusion caused by ordinality. One step further where we drop the first generated value in the binary vector, is known as Dummy Variable Encoding.

The pandas.get_dummies() does One-Hot encoding. To produce an actual dummy encoding from a DataFrame, we need to pass drop_first=True

Using the Dummy Variable Encoding has generated new features as part of the binary vector representation. Hence, now we will reduce the dimensionality of the data.

Dimensionality Reduction:

Variance Inflation Factor (VIF)

When some features are highly correlated, we might have difficulty in distinguishing between their individual effects on the dependent variable. Multicollinearity can be detected using various techniques, one such technique being the **Variance Inflation Factor**. In VIF method, we pick each feature and regress it against all the other features.
Greater VIF denotes greater correlation. Generally, a VIF above 10 indicates a high multicollinearity.

VIF	Features
inf	lead_source_Reference
inf	lead_origin_Lead Add Form
36.408284	lead_origin_Landing Page Submission
33.969414	last_notable_activity_Email Opened
33.624657	last_notable_activity_Modified
28.446795	what_is_your_current_occupation_Unemployed
26.049906	last_notable_activity_SMS Sent
14.806803	lead_source_Olark Chat
14.376434	last_activity_Email Opened

On every iteration we drop the feature with highest vif and check the vif of new features. The final iteration looks like this. For our dataset, we have considered vif threshold of 13

	VIF	Features
12	10.862578	last_activity_SMS Sent
34	10.382051	lead_quality_Not Sure
9	9.735691	last_activity_Email Opened
43	7.858806	last_notable_activity_SMS Sent
30	7.079910	tags_Will revert after reading the email
40	5.465198	last_notable_activity_Modified
31	4.960720	tags_others
0	4.584681	total_visits
29	4.013227	tags_Ringing
22	3.861916	specialization_others
36	3.776454	lead_profile_Potential Lead
2	3.733419	converted
10	3.601342	last_activity_Olark Chat Conversation
11	3.424654	last_activity_Page Visited on Website
16	3.337887	specialization_Finance Management
33	3.322921	lead_quality_Might be
17	3.168489	specialization_Human Resource Management
13	3.156827	last_activity_others
19	3.023177	specialization_Marketing Management
6	2.786412	lead_source_Olark Chat

Sequential Feature Selector (SFS)

Sequential feature selection algorithms are a family of greedy search algorithms that are used to reduce an initial d-dimensional feature space to a k- dimensional feature subspace where k < d.

Recursive Feature Elimination (RFE)

The Recursive Feature Elimination (or RFE) works by recursively removing attributes and building a model on those attributes that remain. It uses the model accuracy to identify which attributes (and combination of attributes) contribute the most to predicting the target attribute.

RandomForestClassifier.feature_importances_

A random forest is a meta estimator that fits several decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the max_samples parameter if bootstrap=True (default), otherwise the whole dataset is used to build each tree.

Machine Learning Models and Evaluation Metrics:

1. **Logistic Regression:** Logistic regression is the go-to method for binary classification problems. It is a supervised learning classification algorithm used to predict the probability of a target variable.

2. **Decision Tree:** The decision tree algorithm tries to solve the problem, by using tree representation. Each internal node of the tree corresponds to an attribute, and each leaf node corresponds to a class label. It a non-parametric supervised learning classification algorithm used to predict class or value of target variables by learning decision rules inferred from prior data.

3. **Naïve Bayes:** It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

4. **Random forest:** An ensemble method is a technique that combines the predictions from multiple machine learning algorithms together to make more accurate predictions than any individual model. A random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting.

5. **Bagging:** Bootstrap Aggregation (or Bagging for short), is a simple and powerful ensemble method used to improve the stability and accuracy of a machine learning algorithm used in statistical classification and regression.

6. **AdaBoost:** Adaptive Boosting (or Adaboost) helps combine multiple "weak classifiers" into a single "strong classifier. It works by iteratively putting more weight on difficult to classify instances and less on those already handled well. Basically, weak models are added sequentially, trained using the weighted training data and thus increases the efficiency of classifiers.

7. **Gradient Boosting:** Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. Unlike Adaboost, this model uses an additive model to add weak learners to minimize the loss function.

Hyperparameter Tuning: It is the problem of choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a parameter whose value is used to control the learning process. In Decision Trees and Random Forest Algorithms, there are many parameters that need to be tuned hence using a cross Fold Validation on a Grid Search provides the best hyperparameters for such machine learning algorithms.

We split the model into a 70:30 ratio. After fitting the engineered data into different models at different iterations, we compared all performances as per below evaluation metrics:

1. **Accuracy:** It determines the overall predicted accuracy of the model.

Accuracy = (TP + TN)/ (TP + TN + FP + FN)

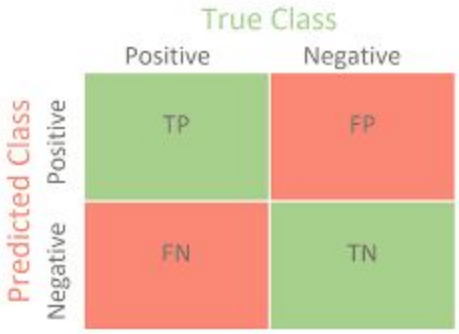
2. **Positive Rate (TPR)/Recall:** It indicates how many positive values, out of all the positive values, have been correctly predicted. It is also known as Sensitivity or Recall.

Recall= (TP/TP + FN)

3. **True Negative Rate (TNR):** It indicates how many negative values, out of all the negative values, have been correctly predicted. It is also known as Specificity.

Specificity = (TN/TN + FP).

4. Precision: It indicates how many values, out of all the predicted positive values, are truly positive. Precision = (TP / TP + FP).



5. F1 Score: F1 score is the harmonic mean of precision and recall. It lies between 0 and 1. Higher the value, better the model.

F1Score = 2*(precision*recall)/ (precision + recall).

6. AUC-ROC: ROC determines the accuracy of a classification model at a user defined threshold value. It determines the model's accuracy using Area Under Curve (AUC). The area under the curve (AUC) represents the performance of the ROC curve. Higher the area, better the model. ROC is plotted between True Positive Rate (Y axis) and False Positive Rate (X Axis). Our aim is to maximize the area under curve. Higher the curve, better the model. At 0.5 threshold, sensitivity = specificity in the ROC Curve

Modeling and Evaluation:

Our model evaluation is done with respect to the best metric for the business problem. Here, it is the recall of the model.

Base Model:

After fitting the data into different models, we have used Logistic Regression for our base model which generated an accuracy of 87%

Model Name	Accuracy Score	Recall Train	Recall Test
LR	87.7	79	74
NB	81.09	75	74
DT	84.8	99	81

Model evaluation using Feature Extraction:

When using new generated feature and modeling the data, we achieved the best results for the Gradient Boosting Model

Model Name	Accuracy	Recall Train	Recall test
LR	86.4	77	78
Gboost	88.8	82	80
RF	87.7	76	75

Model evaluation using Feature Selection by SFS:

On selecting features by Sequential Feature Selection method and modeling the data, we achieved the best results for the Adaptive Boosting Model, with Random Forest as the estimator

Model Name	Accuracy Score	Recall Train	Recall Test
LR	85.5	78	78
NB	77.8	84	83
DT	87.9	79	78
RF	88.09	79	78
Gboost	88.05	79	78
Ada Boost	87.9	79	78

Model evaluation using Feature Selection by RFE:

On selecting features by Recursive Feature Elimination method and modeling the data, we achieved the best results for the Adaptive Boosting Model, with Random Forest as the estimator

Model Name	Accuracy Score	Recall Train	Recall Test
LR	87.7	79	80
NB	81.78	83	84
DT	84.1	96	79
RF	89.5	82	82
Gboost	89.29	82	82
Ada Boost	88.89	92	83

Model evaluation using minimum features by VIF:

On extracting features by reducing dimensionality by VIF method and modeling the data, we achieved the best results for the Adaptive Boosting Model, with Random Forest as the estimator

Model Name	Accuracy Score	Recall Train	Recall Test
LR	87.2	80.6	78.1
DT	85.6	97.2	80.3
NB	81.4	78	77
RF	88.09	79	78
Gboost	89.6	85	83
Ada Boost	89.6	89	83

The recall of the test and train datasets for our AdaBoost Model(RF estimator) is taken from below classification reports:

```
print(metrics.classification_report(ytest,y_predicted))
```

```
              precision    recall  f1-score   support

           0       0.90      0.94      0.92      1682
           1       0.90      0.83      0.86      1090
```

```
print(metrics.classification_report(ytrain,y_predicted_train))
```

```
              precision    recall  f1-score   support

           0       0.94      0.98      0.96      3997
           1       0.96      0.89      0.93      2471
```

As our target is to increase the Converted (1) labels, so we are comparing the recall score of the Converted (1) class to determine our best model.

The best features for converted leads are as below:

	ranks
total_time_spent_on_website	0.253132
lead_quality_Not Sure	0.175391
tags_Will revert after reading the email	0.174654
lead_origin_Lead Add Form	0.120490
last_notable_activity_SMS Sent	0.114439
tags_Ringing	0.054850
lead_quality_Worst	0.033435
tags_others	0.024268
total_visits	0.012954
lead_profile_Potential Lead	0.011885
last_notable_activity_Email Opened	0.008003
lead_source_Google	0.003572
specialization_others	0.002660
lead_profile_others	0.002568
specialization_Finance Management	0.002028
specialization_Marketing Management	0.001813
specialization_Human Resource Management	0.001704
how_did_you_hear_about_x_education_others	0.000864
specialization_Operations Management	0.000836
specialization_IT Projects Management	0.000454

We can interpret that best features are Total_time_spent_on_website, Lead_Quality, Tags, Lead_Origin, Last Notable Activity, Total_Visits, Lead_Profile, Lead_Source, Specialization, How_did_you_hear_about_X_Education

The dummy variable encoding has also clearly provided us the specialization labels that get converted the most:

1. IT Projects Management,
2. Operations Management,
3. Human Resource Management,
4. Marketing Management,
5. Finance Management,
6. Others: which includes Services Excellence, E-Business, Rural and Agribusiness, Retail Management, E-COMMERCE, Hospitality Management, Travel and Tourism, Media and Advertising, International Business, Healthcare Management, specifically people looking to upskill

Summary:

In this project, we aim to analyze the different attributes affecting the hot leads into conversion by building a predictive model and scoring the predicted outcome. We used X education data to perform the experiments necessary. In order to predict the conversion of hot leads, we used the data such as specialization, total_visits, lead_source, occupation, ,lead_quality, lead_origin and other features as input to many machine learning algorithms we gave a try at and selecting the best model.

Feature extraction and feature importance are applied on base models first. Although we have done exhaustive EDA and Feature Engineering to reduce redundant information. After feature engineering and feature extraction pre-processing techniques, hyperparameter tuning is applied to improve the success rates and scalability of the algorithms. Considering the real time usage of the proposed system, achieving better or similar classification performance with minimal subset of features is an important factor for better ML modelling.

For model evaluation we have used a confusion matrix. A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. We have given more importance to Recall score as the significance of False Negative is high i.e., we don't want to miss any lead who is a potential buyer, but our model has classified as negative

After evaluating the base model, we tried to improve the accuracy of our model using feature selection and some complex Algorithms. Feature selection is a process where you automatically select those features in your data that contribute most to the prediction variable or output. The findings show that removing some less important features and owning of business results in a more accurate and scalable system. Therefore, we applied Select KBest, SFS, RFE method.

Cross validation technique is used to check the robustness of a model. Cross Validation is an extremely useful technique for assessing the performance of machine learning models. It is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

Conclusion:

From all models, using cross validation we have selected AdaBoosting Random Forest using combination of
label encoding and get dummies as our best model which gives us better recall score for train and test. AdaBoost is best used to boost the performance of decision trees on binary classification problems.

Our findings support the argument that the features drawn from X education are important for the prediction of target variables

Business Insights:

Key Observations:

1. The more time the user spends on the website, the better their chances of becoming a student.

2. We have observed the probability of a lead being converted into a customer increases with a small decrease of page_views_per_visit from its mean value the pivot point is when the total number of visits on the website is 6. When total visits cross 6, the conversion rate drops as total pageviews decrease.

3. Olark Chat and Direct traffic generate the maximum number of leads. The conversion Rate of Reference leads is high.

4. There exists an 80% conversion rate of working professionals who have previously worked/working in Finance Management, Human Resource Management, and Marketing Management. Unemployed people have been contacted in the highest number, but the conversion rate is low ~40%.

5. The last notable activity being SMS Sent has the highest conversion rate of ~70%. Email opened has a good conversion rate of ~40%.

Recommendation:

1. Approach the leads who spent more time and were not converted and collect the feedback to improve the user experience.

2. For the profiles having current occupation as “Unemployed” the company can look for relevant skills in the lead profile and approach the lead by coming up with offering scholarships and loans.

3. Integration of course recommender into chat bots to suggest to the user the course which interests them the most rather than going through the entire course catalogue and offering some free content in a particular course.

4. Since the ‘Email opened’ last notable activity has a good conversion rate of 40%. The company can go about increasing this interaction rate by reducing the output email for a particular user rather they can go about mentioning free course contents that the user might be interested in.

Future Scope:

For the future scope of the project, we have done Clustering technique using KMeans and Agglomerative Clustering. This is so that we get a clearer understanding of the actual labels instead of already provided target labels and compare it to the supervised technique.

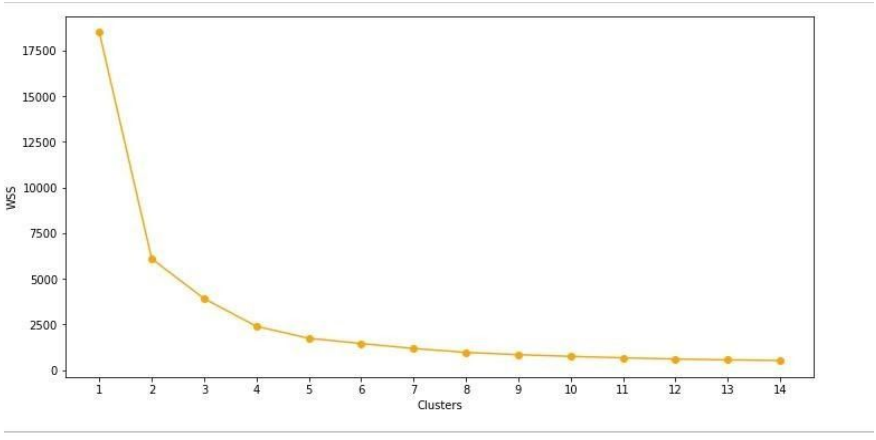
Semi-supervised clustering is a technique that partitions unlabelled data by making use of domain knowledge, usually expressed as pairwise constraints among instances or just as an additional set of labelled instances. We have used k-means clustering to cluster our data. k-means is one of the simplest unsupervised learning algorithms that solves the well -known clustering problem.

The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters). The main idea is to define k centres, one for each cluster. These centres should be placed in a cunning way because different locations cause different results. So, the better choice is to place them as much as possible far away from each other.

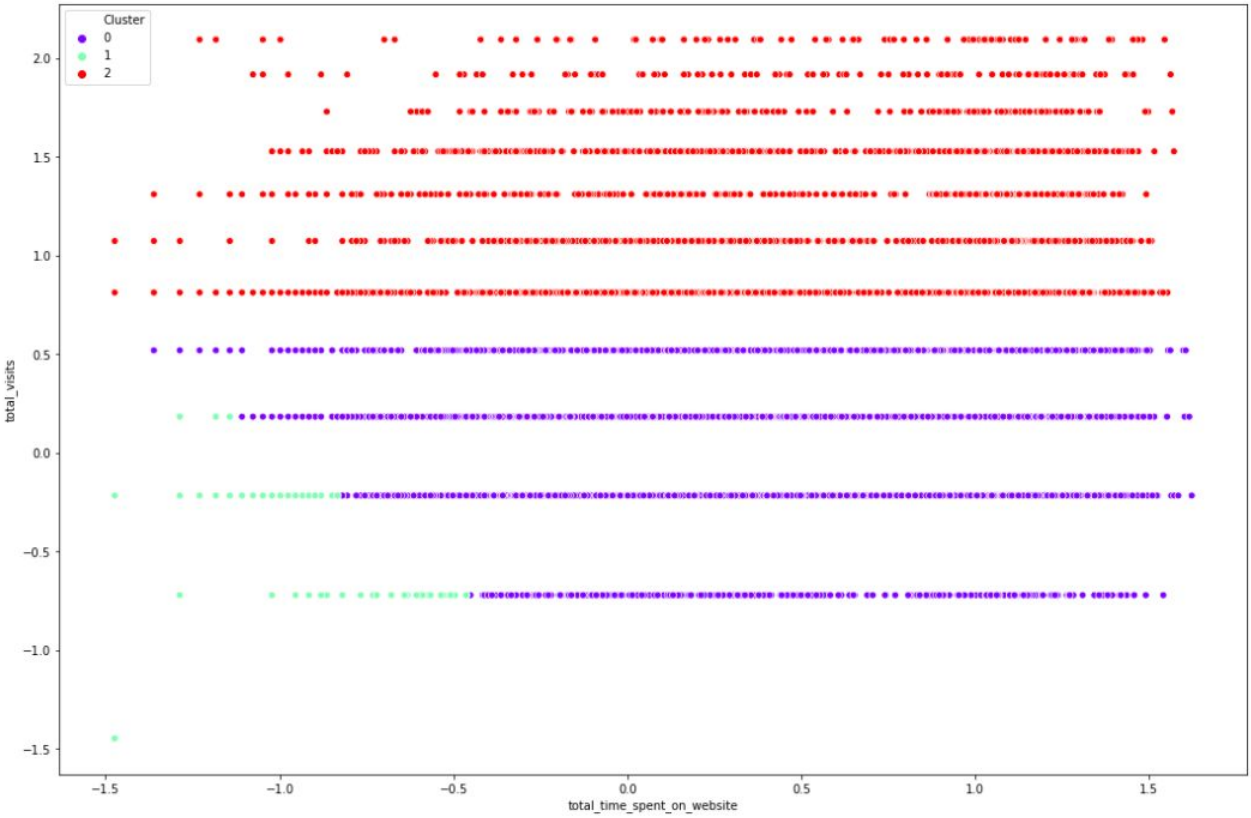
Each cluster provides a WSS value (within-**cluster** sum of square). **WSS** means the sum of distances between the points and the corresponding centroids for each **cluster.** For our dataset, we have below WSS value for different clusters:

num of clusters	wss_value
1	18480.000000
2	6099.307245
3	3919.068493
4	2400.308346
5	1746.754537
6	1457.477857

We determine the optimal number of clusters from the elbow plot of the cluster wss values



The elbow point is 3, hence we use 3 clusters in our target label and predict using KMeans Clustering. We can see the clusters using any two features and passing the new labels as hue



From the understanding of the dataset we can say that the data holds more value if we keep 3 labels- Converted, Not Converted and Could be Converted. On basis of the best features X Education can get most conversion from our 0th label Could be Converted and work on those hot leads.

When we use these new labels for multiclass model, we see extremely good models coming along

Model Name	Accuracy Score	Recall Train	Recall Test
LR	99.7	100	100
RF	99.4	100	100

Here we see that we get good precision and recall scores for both Logistic & Random Forest Models. We shall go ahead with the LR model for multiclass target labels. Here Class 2- Converted, Class 1- Not Converted and Class 0- Could be Converted. Hence, correctly predicting Class 0 customers is helpful for X Education's increase of Conversion rate as they should be focused more on these leads.

The LR model gives no error in Class 1 and Class 2 and it gives lesser error than Random Forest Model for Class-0 as it is the most important class to focus on to increase sales.

We can also collect real time data from institutes like X Education to analyse and increase the efficiency of our unsupervised model.

Reference documents:

- https://www.kaggle.com/amritachatterjee09/lead-scoring-dataset
- https://www.geeksforgeeks.org/deploy-machine-learning-model-using-flask/
- https://scholarspace.manoa.hawaii.edu/bitstream/10125/63916/1/0143.pdf
- https://objectiveit.com/blog/use-ai-and-machine-learning-for-predictive-lead-scoring/