Logistic Regression Case Study

LEAD CONVERSION ANALYSIS

DATE:20-MAY-2024

TEAM MEMBERS: SAYEDA MUBASHRA ALI, NAGARJUN R, MEGHANA SHETTY

Executive Summary

Problem Statement

Industry-Education : X Education sells online courses to professionals, acquiring leads through their website, Google, and referrals.

- Interested visitors/Leads can browse courses, fill out forms
- The sales team engages these leads via calls and emails, but only about 30% convert into customers.The low conversion rate highlights inefficiency, with only 30 out of 100 daily leads typically converting.

Objectives

To improve efficiency, the company aims to identify 'Hot Leads' to focus sales efforts on high-potential leads, increasing the conversion rate

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 (higher score means higher conversion possibility and vice versa)
2. Possible solutions to additional problems presented by the company
3. CEO has given a target lead conversion rate to be around 80%, we aim to build a model with a sensitivity >=80%

APPROACH

- Understanding the data/variables
- Import the data
- Identifying the data types
- Review the headers and content of the files

- Univariate analysis
- Bivariate Analysis
- Multivariate Analysis

- Evaluate the model's performance using metrics such as accuracy, precision, recall, and F1 score.

- Imputing/removing missing values
- Handling outliers
- Standardising the values
- Fixing invalid values

- Identify key features that influence lead conversion
- Develop and train a predictive model to assign lead scores based on the likelihood of conversion.
- Experiment using feature Engineering

- Identifying top features that influence the target variable
- Recommendations for decision making

Data Cleaning

Missing Values

1. Many of the categorical variables have a level called 'Select' which needs to be converted to null value.The drop down may have had select as the default value which was not changed while capturing the data ('Specialization', 'How did you hear about X Education', 'Lead Profile' & 'City')

2. There are few columns with missing values accounting for more than 40%, in such cases the columns were removed ('How did you hear about X Education', 'Lead Profile', 'Lead Quality', 'Asymmetrique Profile Score', 'Asymmetrique Activity Score', 'Asymmetrique Activity Index', 'Asymmetrique Profile Index',City)

3. Imputing values for missing data in Specialization column as a separate value- Other (36%).

4. Dropping the few columns as they don't add valuable information for modelling or because there is dominance of just 1 value other than missing values (Tags, 'What matters most to you in choosing a course', Country, 'Prospect ID', 'Lead Number')

5. Imputing missing values for columns like'TotalVisits', 'Page Views Per Visit', 'Last Activity', 'Lead Source' :replacing these with the most frequent value

Data Cleaning

Standardizing values

1. Dropping few columns which has only 1 value and hence not useful in modelling(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'

2. Grouping low frequency value in columns to Others(Lead Source, Last Activity')

3. Dropping columns which have dominance of more 90% of 1 value as it may be skew the model results ('Do Not Call','Search','Newspaper Article','X Education Forums','Newspaper','Digital Advertisement','Through Recommendations')

4. Flags were converted to binary values for further analysis ('Do Not Email', 'A free copy of Mastering The Interview')

Outliers Identification and treatment

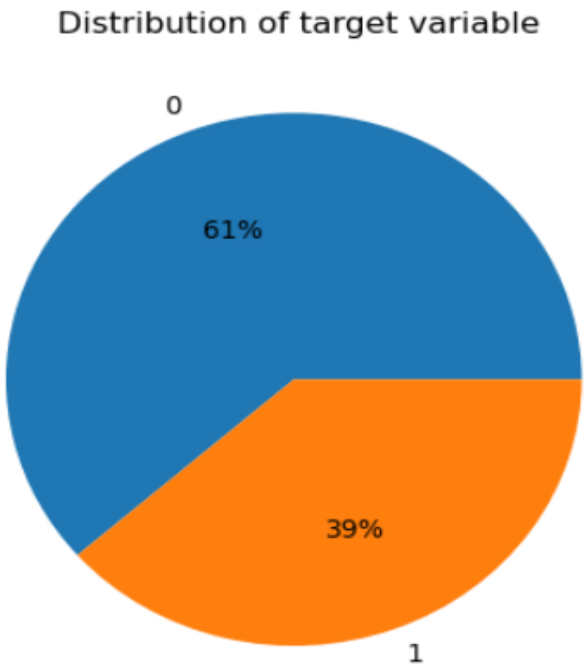
1. Removed outliers more than 1.5 times the IQR (Interquartile Range) to enhance data integrity and model performance for two columns- TotalVisits", "Page Views Per Visit"

2. Post the application of the function, we hope to achieve improvements in model accuracy and reliability, especially for sensitive models like logistic regression.

3. We also see that this provides clearer data visualizations by eliminating distortion caused by outliers.

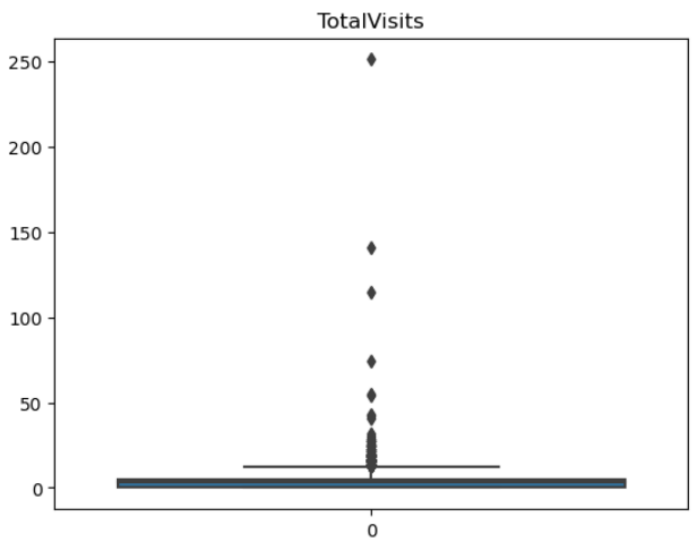
Data Imbalance

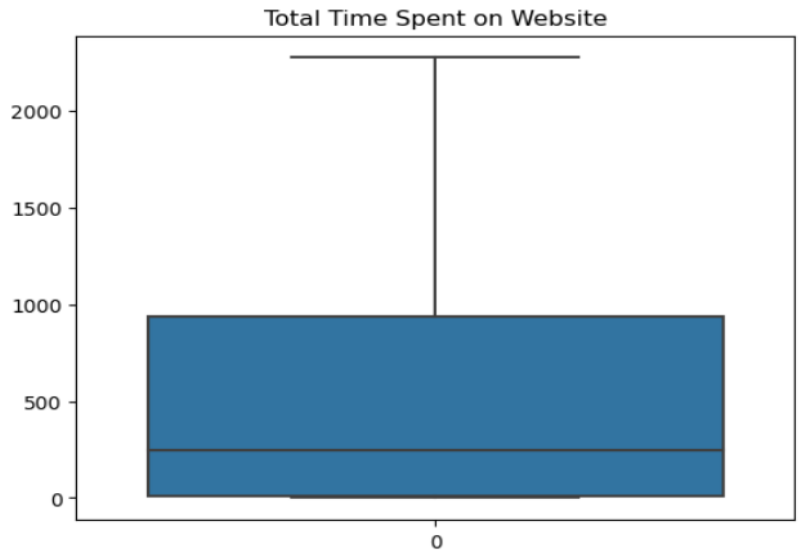
Target column- 1 Converted and 0 indicates leads that are not converted

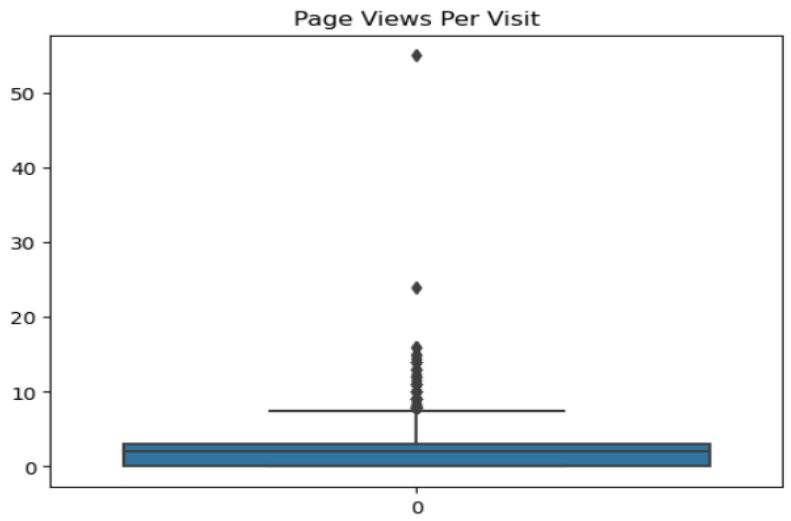


Of the total dataset, 39% of the data belongs to converted leads, looks like we have **slight imbalance in the data**

Univariate Analysis- Numerical

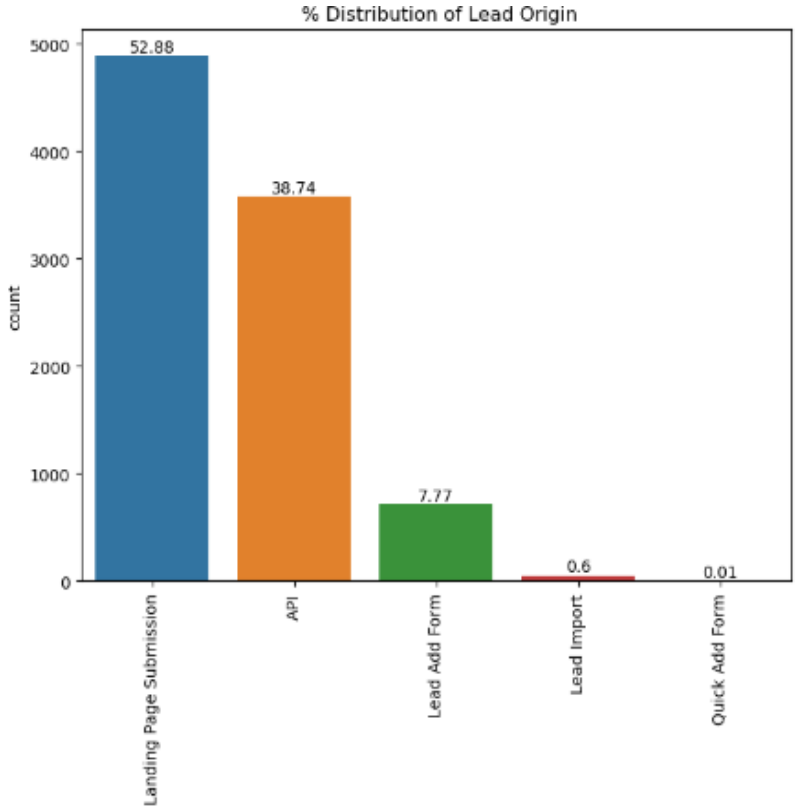


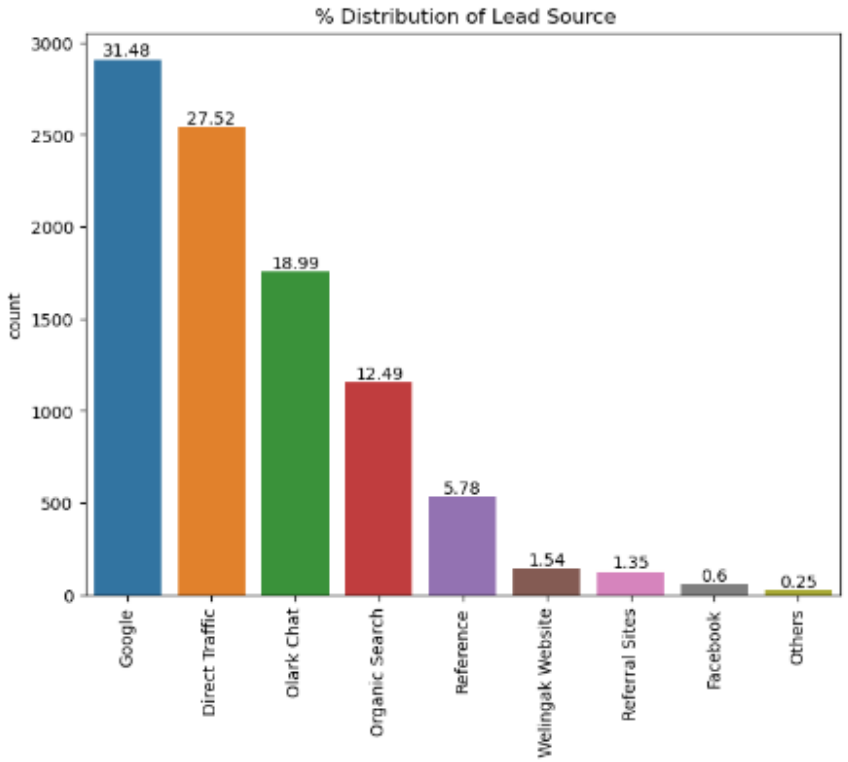




- Outliers seen in TotalVisits“ and "Page Views Per Visit“, we ‘ll cap the outliers outselide the range of upper and lower bounds i.e +-1.5 IQR with the value at upper and lower bound

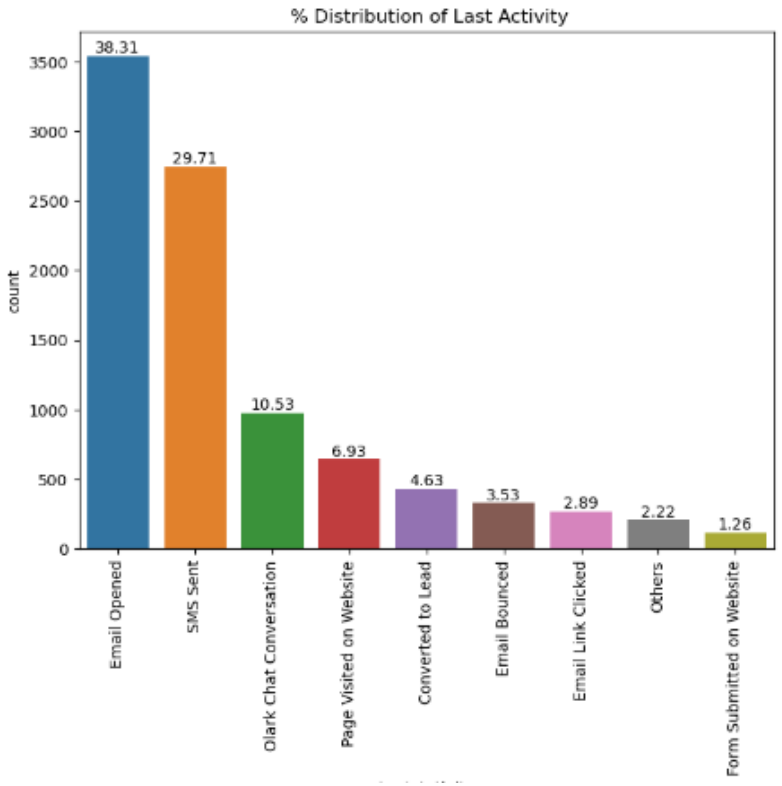
Univariate Analysis- Categorical(1)

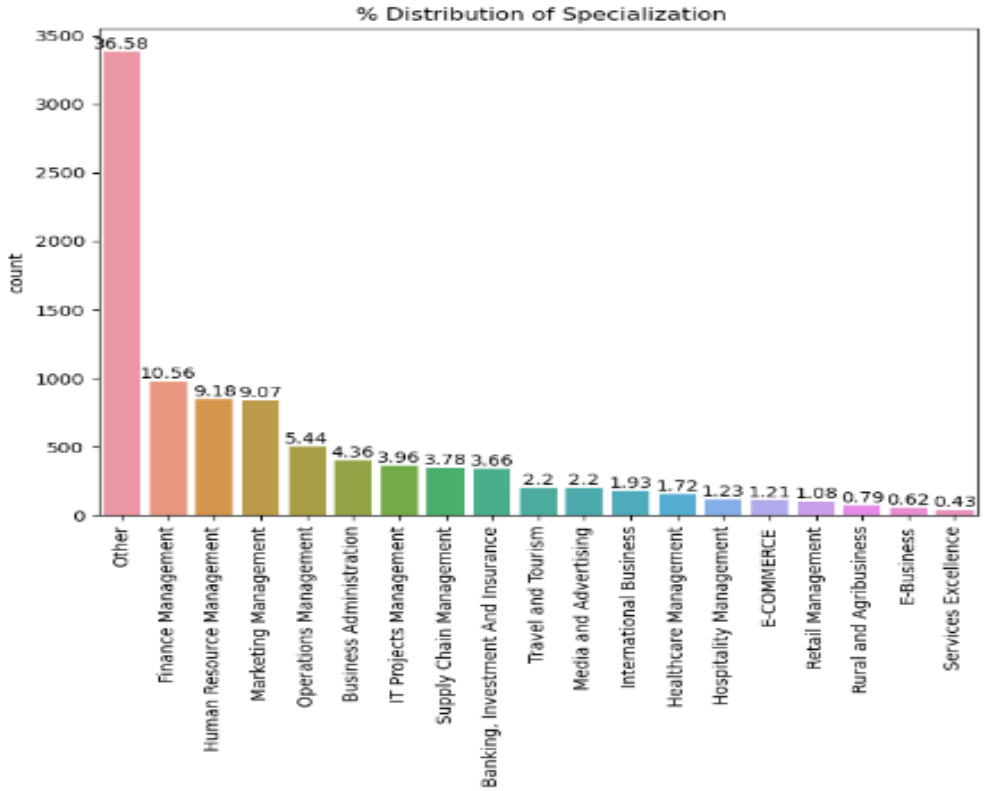




1.Lead Origin: Highest number of leads were identified though "Landing Page Submission" closely followed by API , accounts for more than 80%
2.Lead Source- Approx 60% of the lead source is from Google and Direct Traffic

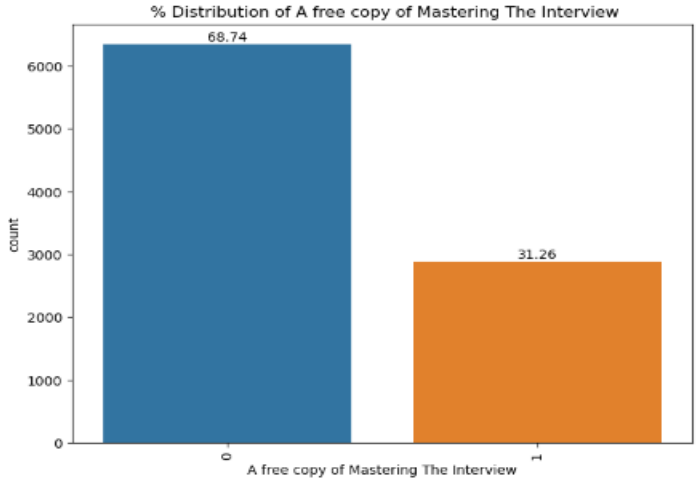
Univariate Analysis- Categorical(2)

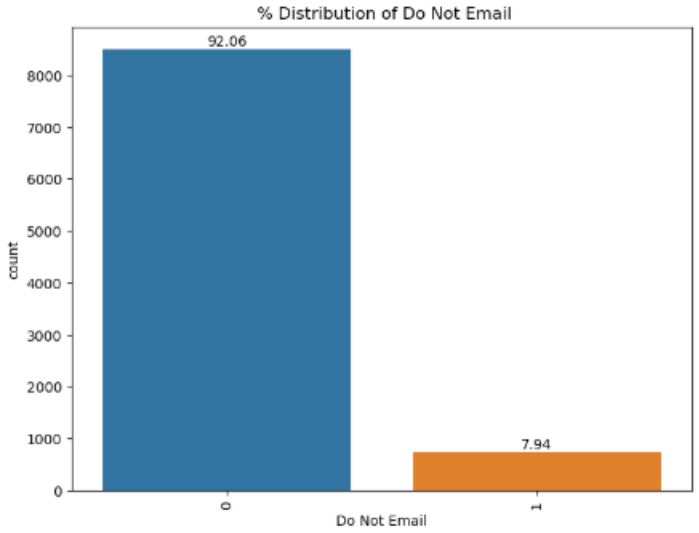


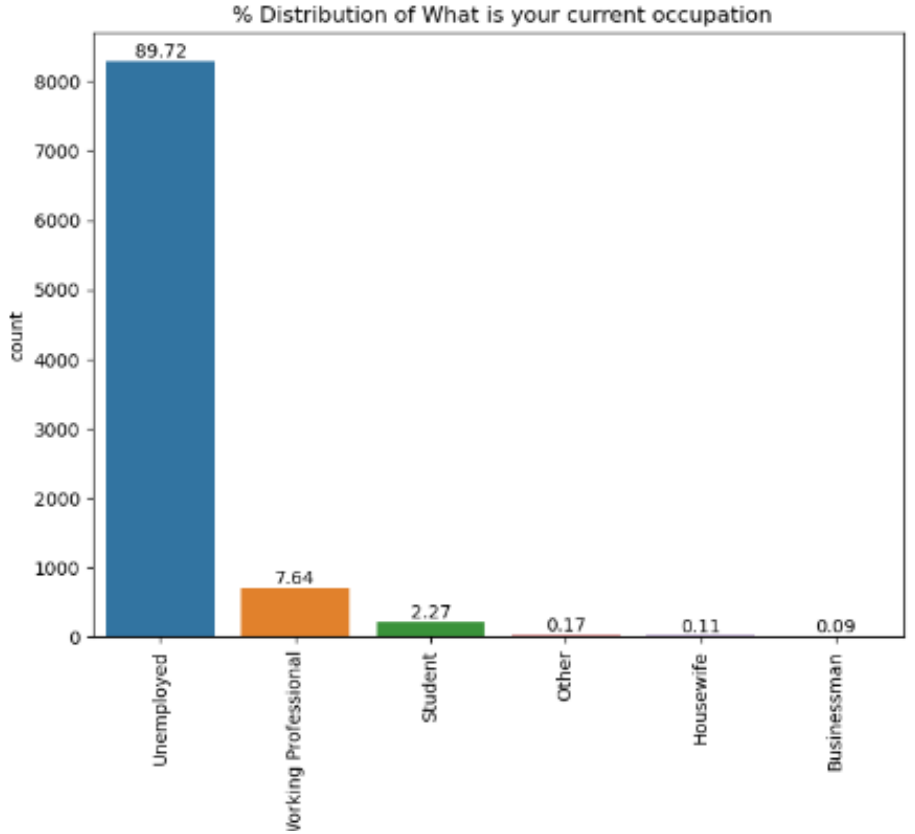


1. Last Activity- Approx 70% of customers last activity is captured as SMS Sent & Email Opened

Univariate Analysis- Categorical(3)

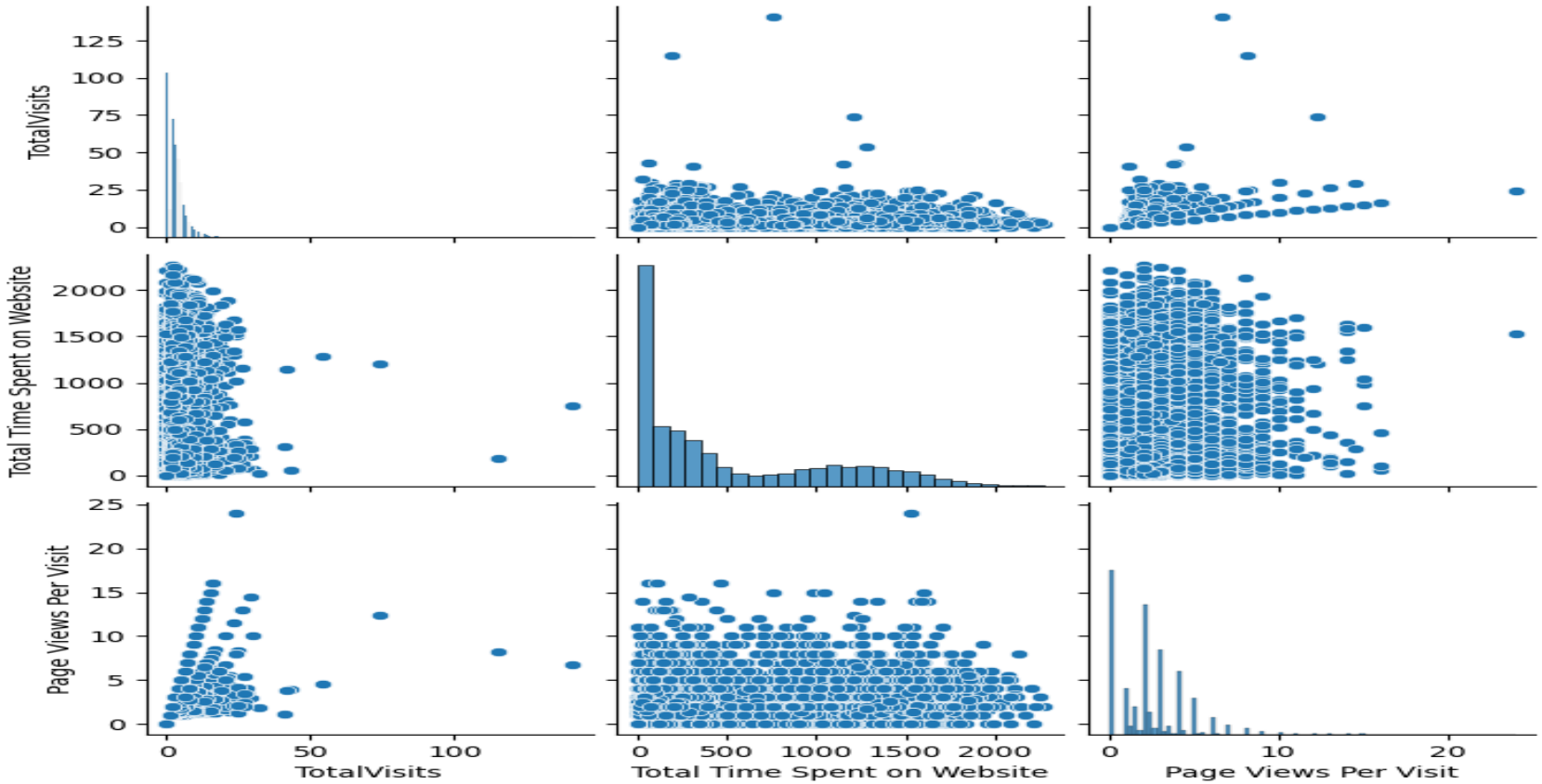






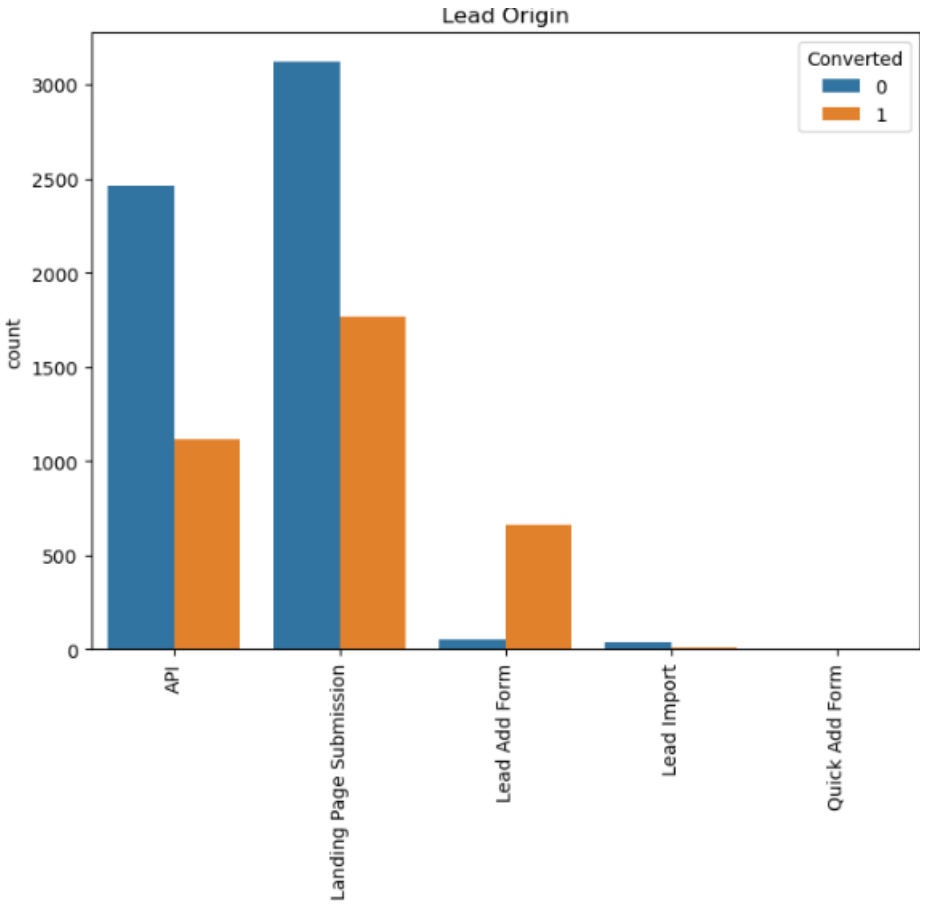
1. Lead Origin: Highest number of leads were Occupation- 90% of the customers are unemployed
2. More than 90% of customers have indicated no in the do not email field

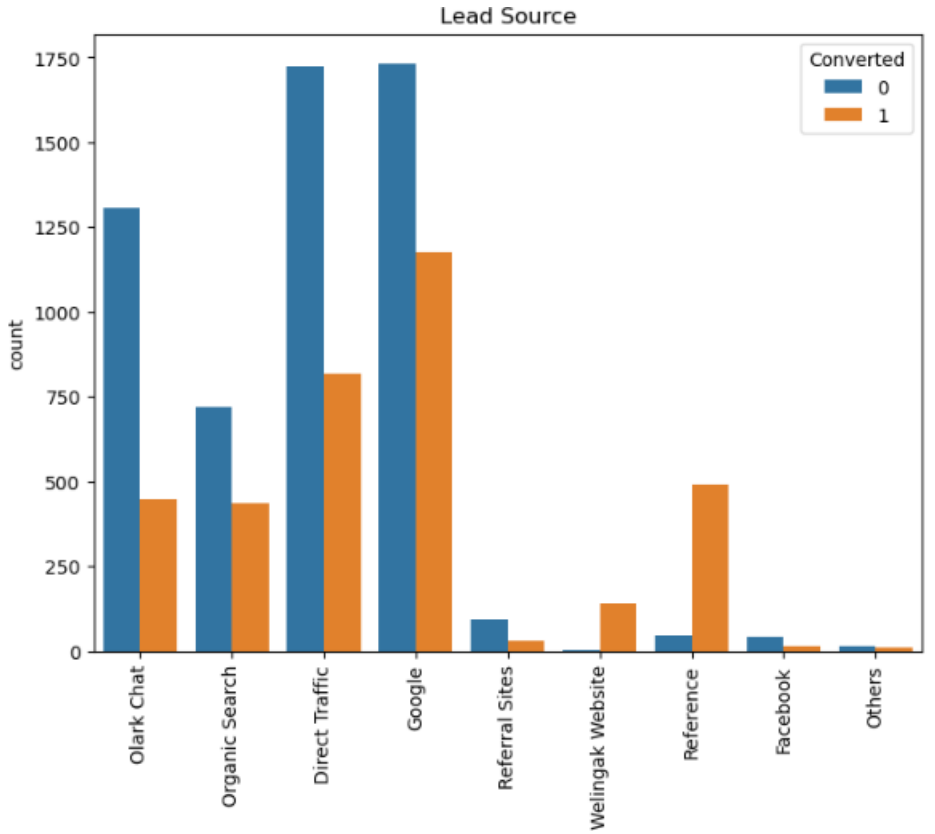
Bivariate Analysis(1)



There is a some positive correlation Page views per visit and Total Visits

Bivariate Analysis(2)

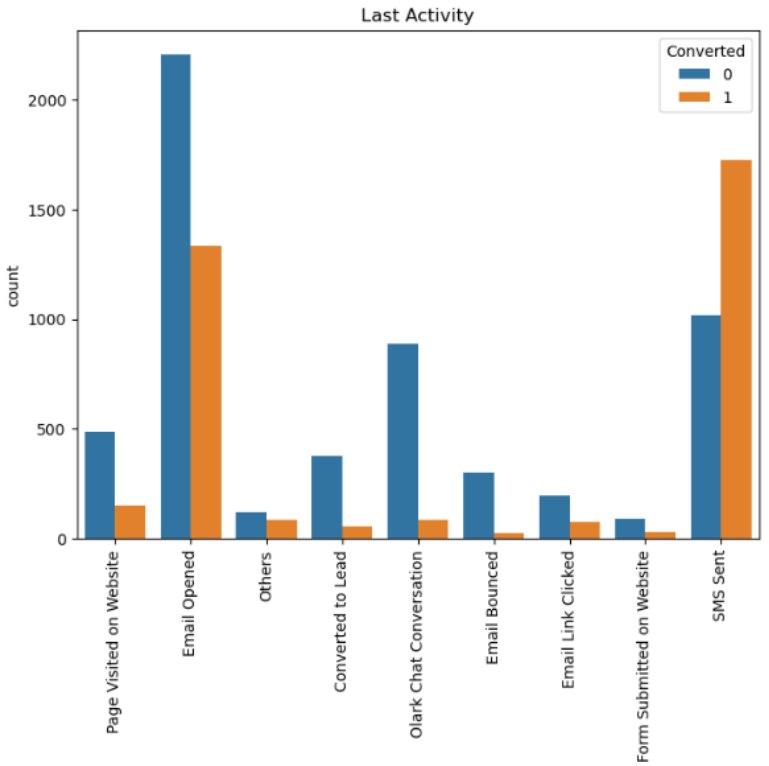


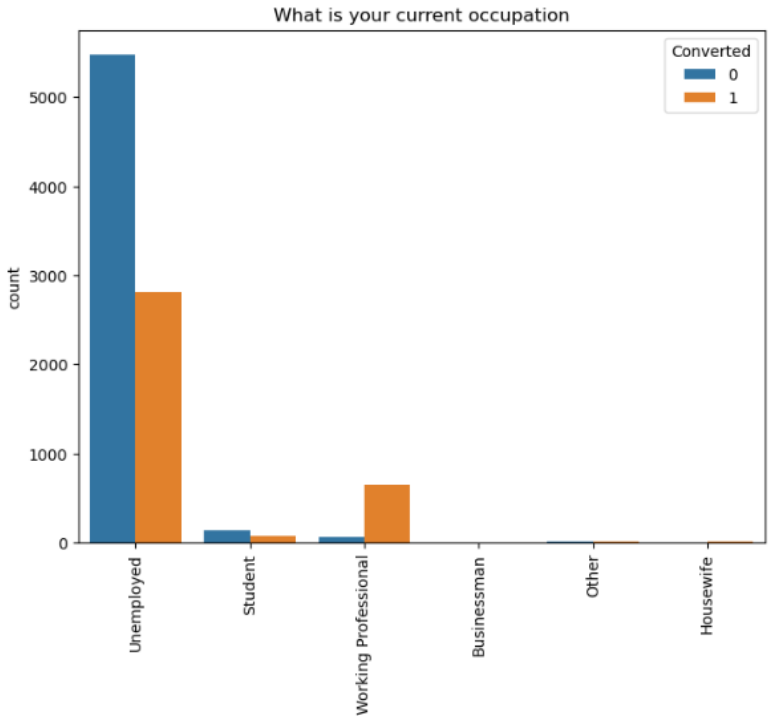


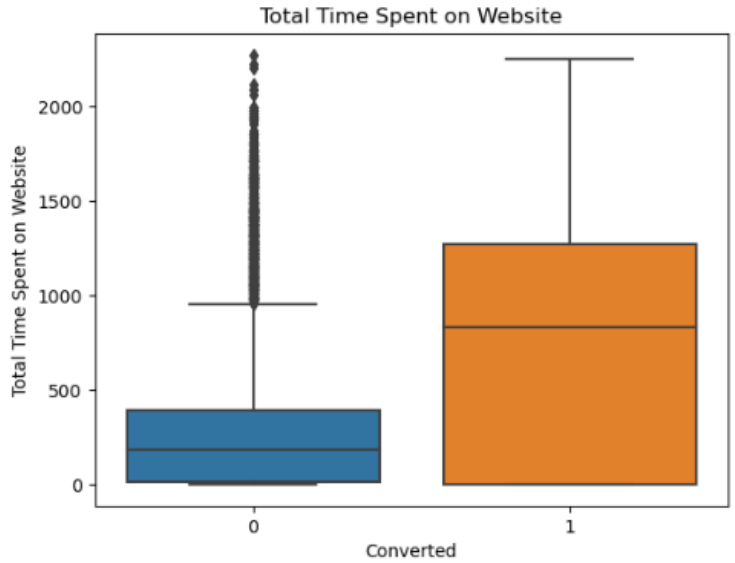
1.Lead Origin: Conversion rate is higher when the origin is Landing page submission or API, the converstion ratio is higher when origin is lead add form

2.Lead Source: the conversion rate is higher when the source is reference akthough teh total customers from this is lower than Google and Direct Traffic

Bivariate Analysis(3)





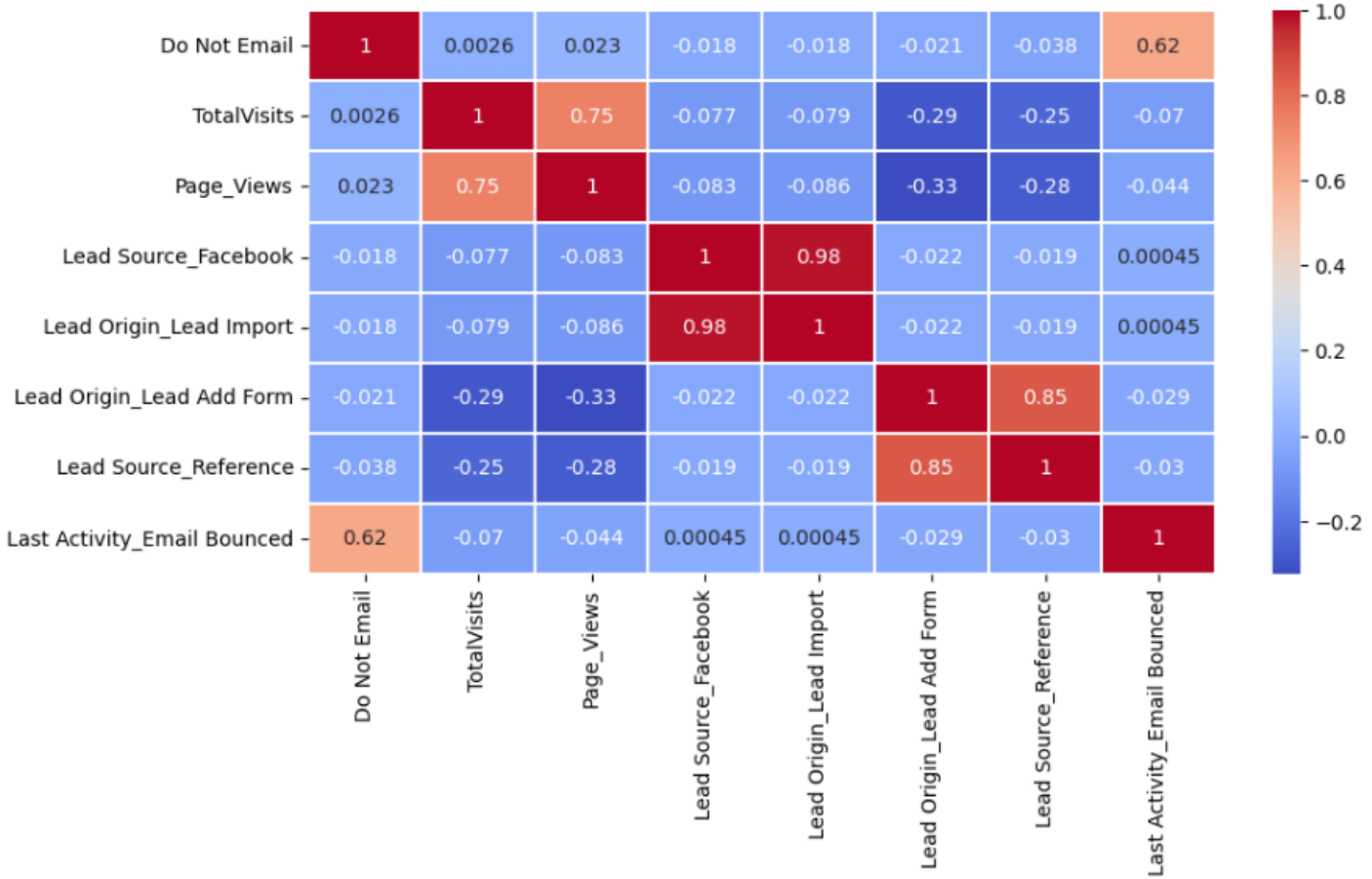


1.Occupation: Lead conversion rate is higher in Unemployed and Working Professional although total customers in the working prof category is low

2.Last Activity: 'SMS Sent' has high lead conversion rate followed by Email Opened', almost 80%

3.Conversion rate is higher when the time spent is higher 250 as compared to 1000

Multivariate Analysis



Correlation between columns

1. Page view visit is correlated with Total visits i.e 0.75
2. There is high correlation between source refernec and origin addform
3. Email bounced and do not email has a high correlation 0.62

Modelling

Data Preparation for Modelling

1. Creating a dummy variable for categorical variables and dropping the first one("Lead Origin","Lead Source","Last Activity","Specialization","What is your current occupation“)
2. Data Splitting: Training Set: 70%, Test Set: 30%
3. Feature Scaling: Used StandardScaler for numerical features.

Modelling

1. Feature Selection: Recursive Feature Elimination (RFE) to select the top 15 features.
2. Multicollinearity Check by calculating Variance Inflation Factor (VIF) to addressed high VIF values(>5)
3. Created another model after dropping feature 'What is your current occupation_Housewife' as p value is high

Model Evaluation

Evaluation Metrics

1. Predicted probability was calculated using the final model and created roc curve

2. roc curve is used to evaluate the performance of binary classification models by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR). The Area Under the Curve (AUC) summarizes the model's performance: AUC = 1: Perfect model, AUC = 0.5: Random guess, AUC < 0.5: Worse than random. AUC for our model is 0.88 which is closer to a perfect model

3. Probability cut off of 0.35 was used to predict conversion after plotting accuracy, specificity and sensitivity values for different probabilities and choosing the best value

4. Precision-recall curve is a trade-off between precision and recall at different thresholds. A higher area under the curve (AUC) indicates better model performance. Evaluated the rest of the metrics using probability cut off as 0.41 and saw the fall in the model performance. Finally cutoff value was reverted to 0.35

5. Model is Stable as the train set metrics and Test metrics are comparable and close to each other. Sensitivity is ~81% when the probably cutoff is chosen as 0.35. Sensitivity indicates converted leads identified by the model correctly of all potential leads. This is inline with the target set by CEO

Training Model Metrics
True Negatives: 3243 True Positives: 1989 False Positive: 759 False Negative: 477
Accuracy: 81% Sensitivity/Recall: 81% Specificity:81% Precision:72% F1 Score: 76%

Test Model Metrics
True Negatives: 1373 True Positives: 878 False Positive: 304 False Negative: 217
Accuracy: 81% Sensitivity/Recall: 80% Specificity: 81% Precision: 74% F1 Score: 76%

Model Params and Coefficients

1. Lead Origin_Lead Add Form 3.046808
2. What is your current occupation_Working Professional 2.644358
3. Lead Source_Welingak Website 2.561375
4. Last Activity_SMS Sent 1.909112
5. Last Activity_Others 1.392287
6. Total Time Spent on Website 1.066049
7. Lead Source_Olark Chat 1.022836
8. Last Activity_Email Opened 0.751451
9. Last Activity_Olark Chat Conversation -0.703482
10. const -0.952024
11. Specialization_Hospitality Management -1.065775
12. Lead Origin_Landing Page Submission -1.124447
13. Specialization_Other -1.182373
14. Do Not Email -1.216202
15. dtype: float64

1.Model has 13 features, Top 3 features- Lead Origin_Lead Add Form and What is your current occupation_Working Professional, Lead Source_Welingak Website

2.Sign of the coefficients indicate the direction of the relationship with the model outcome i.e Leads originating from landing page submissions are less likely to convert (odds ratio:0.31) and Leads from the lead add form are significantly more likely to convert (Odds Ratio: 16.34)

Recommendations

1. Enhance Engagement for Working Professionals: As working professionals are significantly more likely to convert, tailor marketing messages and campaigns specifically to this demographic. Offer content and programs that cater to the career advancement needs of working professionals.

2. Leverage Effective Lead Sources/Lead origins: Leads from lead add forms have a high conversion rate, ensure that these forms are easily accessible and prominently featured on your website.

3. Promote Welingak Website: Increase traffic to the Welingak website, as it is a highly effective lead source. Consider paid ads, and partnerships to drive more visitors.

4. Utilize SMS Effectively: Leads who received SMS messages are significantly more likely to convert. Develop and implement targeted SMS marketing campaigns to increase engagement and conversions. Use SMS for timely follow-ups after important actions (e.g., form submissions, content downloads) to keep the leads engaged.

5. Increase Website Visit Time: Since more time spent on the website correlates with higher conversion rates, enhance website content and user experience to keep visitors engaged longer. Monitor pages most visited and optimize them to improve user engagement and conversion rates.

6. Diversify Offerings: Certain specializations have lower conversion rates, consider diversifying your offerings to appeal more broadly

By focusing on the above, you can improve lead conversion rates and optimize marketing efforts. These recommendations are based on the significant predictors identified in the logistic regression model, ensuring that marketing strategies are data-driven and targeted for maximum effectiveness. To maximize lead conversion during the internship period, the sales team should focus on prioritizing their outreach efforts to potential leads who have been predicted by the model to convert.

Thank you