1. Introduction

Company named X Education gets a lot of leads. However, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone. 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.

2. Business Problem

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.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these 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%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

3. Data Description

In this project, we are provided with two datasets: "Leads.csv" and "Test.csv". These datasets will be used for performing exploratory data analysis (EDA) and building a logistic regression model.

a. Leads.csv Dataset:

- This dataset contains 37 columns and 9240 rows.

- It is the main dataset on which we will perform EDA and build our logistic regression model.

- The dataset includes various features or variables related to leads or potential customers, along with a target variable indicating whether the lead got converted or not.

- The columns contain information such as lead demographics, source of leads, marketing interactions, website activity, communication history, and more.

- I explored the data, analyzed the relationships between variables, handle missing values, perform feature engineering if necessary, and pre-process the data for model building.

- The target variable(Converted) will be used for training the logistic regression model to predict the likelihood of lead conversion.

b. Test.csv Dataset:

- This dataset contains 36 columns and 2007 rows.

- It is a separate dataset provided for testing the logistic regression model prepared using the "Leads.csv" dataset.

- The purpose of this dataset is to evaluate the performance of the trained model on unseen data.

- It contains similar columns as the "Leads.csv" dataset, except for the target variable which is not included in this dataset.

- We will use the trained logistic regression model to predict the likelihood of lead conversion for the leads in this dataset and evaluate the model's performance.

Overall, this project involves analyzing the "Leads.csv" dataset, performing EDA to gain insights into the data, building a logistic regression model using the provided features, and then testing the model's performance on the separate "Test.csv" dataset. The goal is to develop an accurate and reliable model that can predict the probability of lead conversion based on the available information.

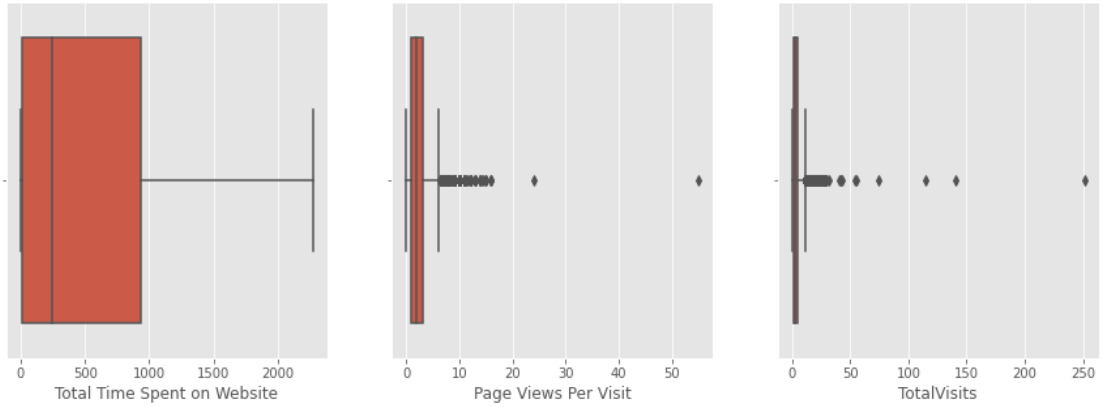
4. Data pre-processing

Data pre-processing is an essential step in preparing the "Leads.csv" dataset for exploratory data analysis (EDA) and building a logistic regression model. It involves handling missing values, handling categorical variables, scaling numerical variables, and performing any necessary feature engineering. Here's an outline of the data pre-processing steps done for the "Leads.csv" and “test.csv” datasets:

Handling Missing Values: A lot of columns had more than 40% of Null values, these columns were dropped. For columns with less than 40% NaN values further investigation and imputation was carried out.

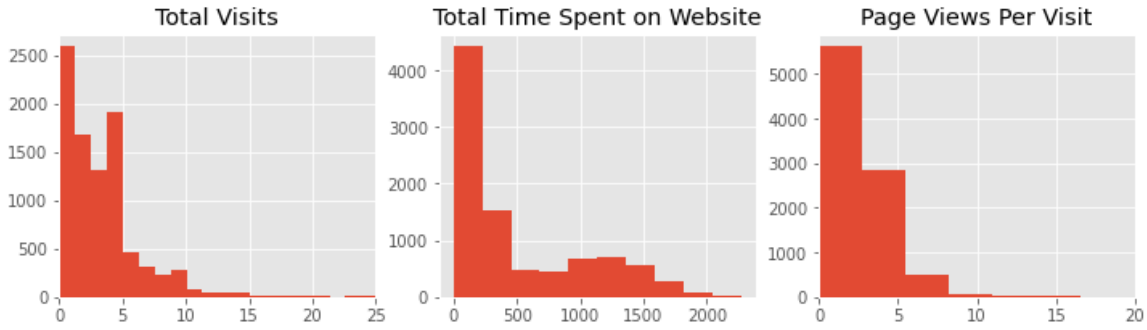
Segmentation: Columns were divided into categorical and numerical variables and further investigation was carried out on them accordingly.

Handling Outliers: The continuous columns were checked using box plot for outliers. There were some outliers which were imputed using capping method.

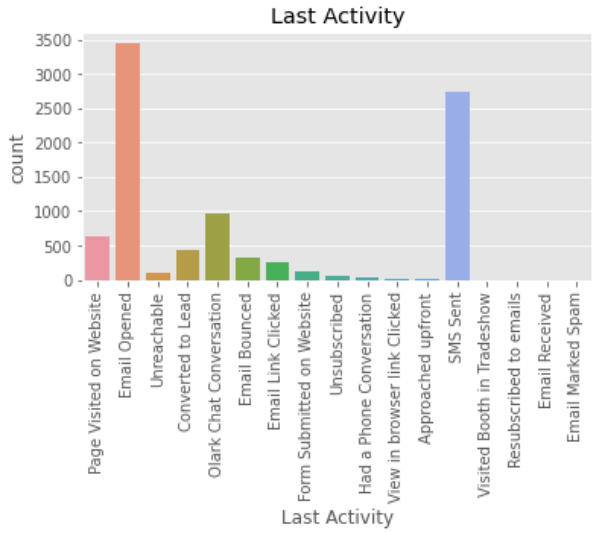


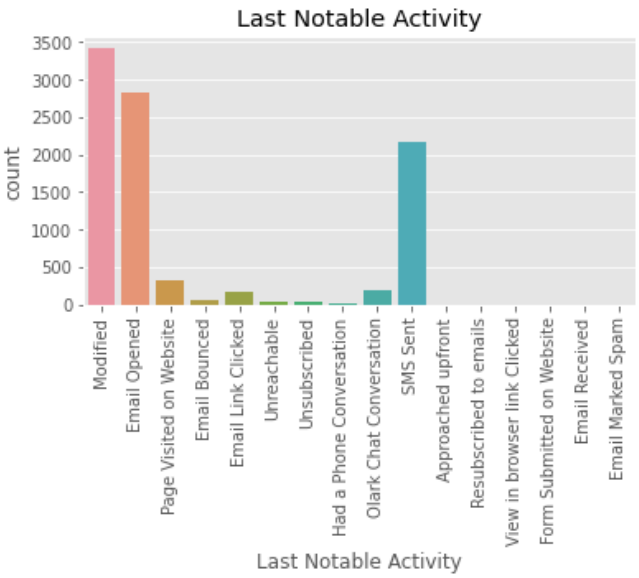
5. EDA

Univariate Analysis of the dataset was carried out to gain better understanding of the variables.



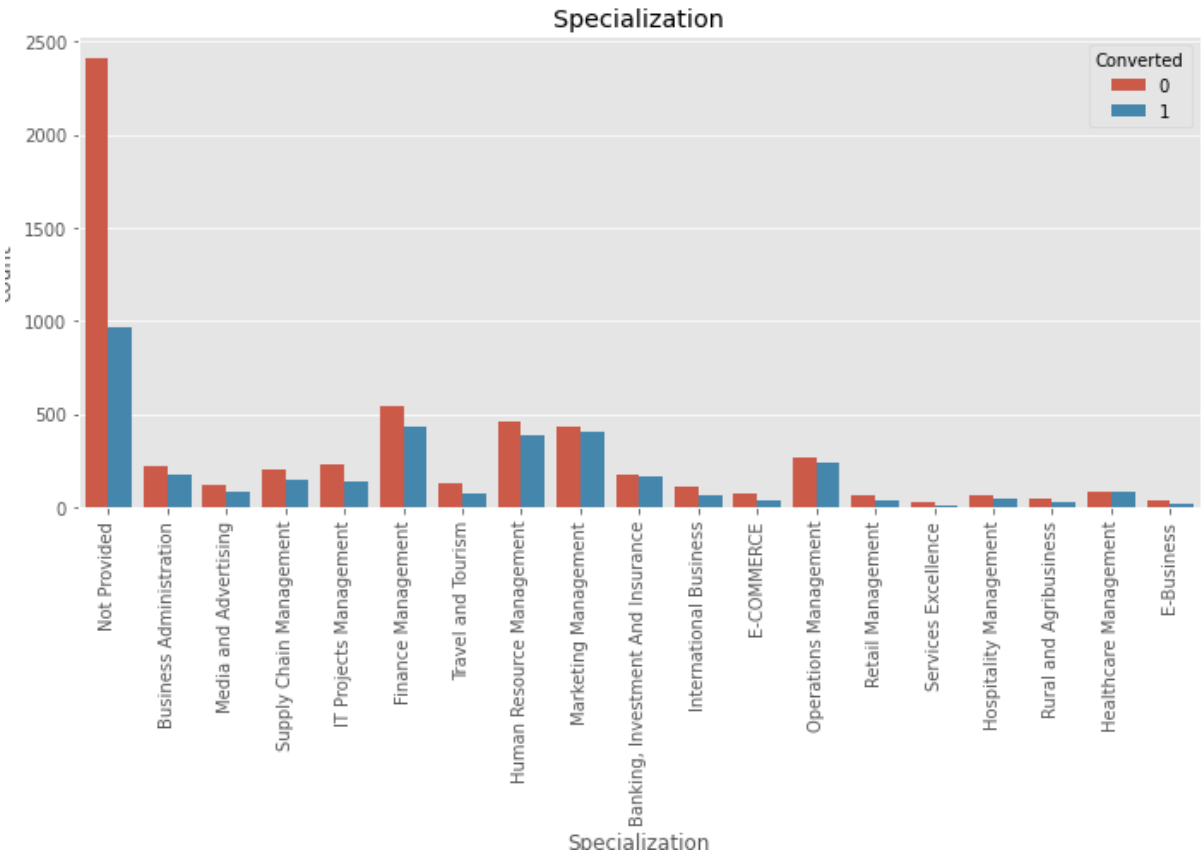
The Continuous columns suggested the data was skewed to the left.

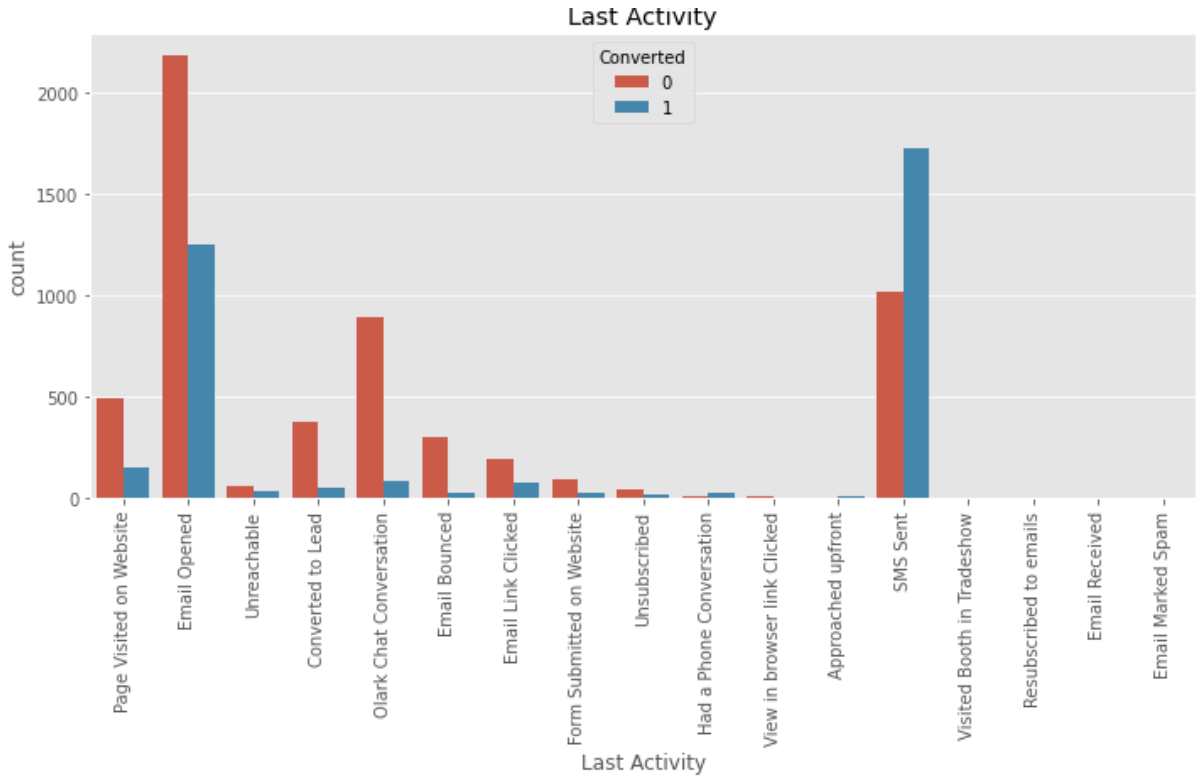




These features seem to have significant influence on the Hot Leads prediction.

Bivariate Analysis





These columns give insights into the matter showing certain significant factors that might affect the model.

6. Model Development

Model Development involves building a logistic regression model using the pre-processed dataset ("Leads.csv") to predict the probability of lead conversion. Here's an outline of the steps involved in model development:

Dummy Variables – Dummy variables were created for features like 'Lead Origin', 'Lead Source', 'Last Activity', 'Specialization', 'What is your current occupation', 'Tags', 'City', 'Last Notable Activity'. To incorporate these values to the model.

Train-Test Split - The dataset was split into training dataset and testing dataset. This was performed in order to train the model from the same dataset.

Scaling – Standard scaler was used for scaling the training and testing data.

Model - Logistic regression model with StatsModels was used for making the model.

Firstly, RFE was used to limit the features to 15 features in count.

Constant was added to the set to get the accurate values of the features coefficients.

Multiple iterations of VIF calculation followed by dropping of improper features were performed to get best results.

7. Model Evaluation

Model was evaluated on basis of certain metrics:

For model with arbitrary cut-off 0.5

Confusion Matrix:

[[3549 , 154]

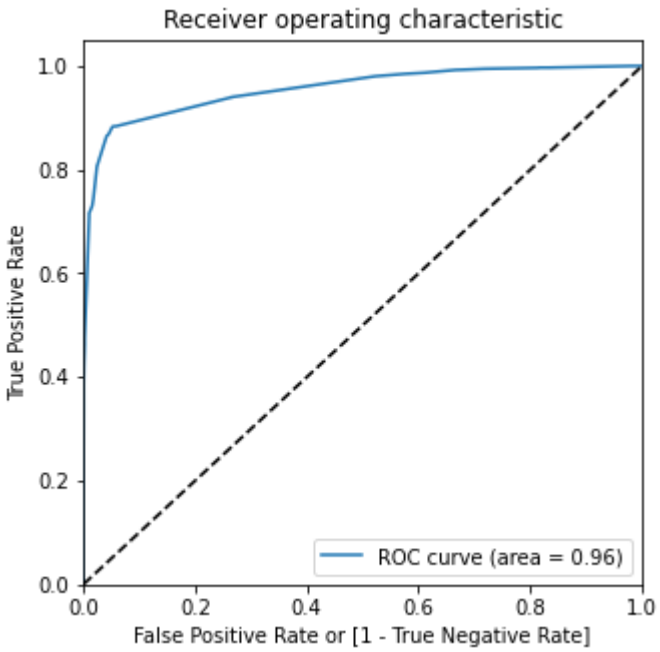
[302, 1926]]

The model has an accuracy on 90% nearly.

Other metrics –

- sensitivity: 0.86
- specificity: 0.95
- false_postive_rate: 0.041
- positive_predictive_value: 0.92
- negative_predictive_value: 0.92

ROC Curve



Optimal cutoff probability is where we get balanced sensitivity and specificity.

8. Model Testing

The final stage of model testing was done on 'Test.csv'. First, the dataset was cleaned of missing values and outliers.

Dummy variables were created and the model was applied on the test data.

The output result csv was uploaded on Kaggle.

The final Kaggle score achieved was approximately 0.91524.

9. Final Suggestions

- Focus on lead nurturing activities: Prioritize personalized emails, SMS messages, and targeted newsletters to maintain engagement and increase conversion chances.
- Utilize automated SMS messages: Target leads with high conversion potential by sending them personalized SMS messages using the predictive model.
- Collaborate with relevant teams: Work closely with management, data scientists, and other teams to fine-tune the model and gather feedback for continuous improvement.
- Offer discounts or incentives: Develop strategies to provide discounts or incentives that create urgency and motivate leads to make a purchase decision.
- Diversify communication channels: Engage with leads through channels like email, social media, or chatbots to reduce reliance on phone calls.
- Gather customer feedback: Seek feedback from existing customers to refine lead generation strategies and improve the conversion process.