Optimizing Lead Conversion: Unleashing Hot Leads with Effective Lead Scoring

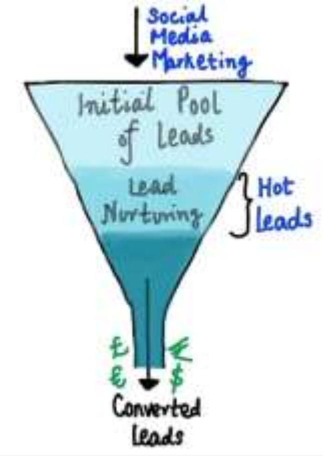
By:-
Mohammed Faiz
Mohammed Mustafa C
Afridi Sheik

Problem Statement

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

The company markets its courses on several websites, search engines, and even social media sometimes. Once these people land on the website, they might browse the courses, fill out a form for the course, or watch some videos. When these people fill out a form with their email address or phone number, they are classified as leads. 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 into successful sales, while most of the leads do not. The typical lead to successful sale conversion rate at X education is around 30%.

Now, although X Education gets a lot of leads, its lead-to-sale conversion rate is very poor. For example, if they acquire 100 leads in a day, only about 30 of them are converted into successful sales. 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 would go up as the sales team would now be focusing more on communicating with the potential leads rather than making calls to everyone. A typical lead conversion process can be represented using the following funnel.



As you can see, there are a lot of leads generated in the initial stage (the initial pool of leads), but only a few of them come out as paying customers from the bottom (converted leads). In the middle stage (lead nurturing), you need to nurture the potential leads well (i.e., educate the leads about the product, constantly communicate, etc.) in order to get a higher lead conversion.

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 a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark estimate of the target lead conversion rate as being around 80%.

Data Given

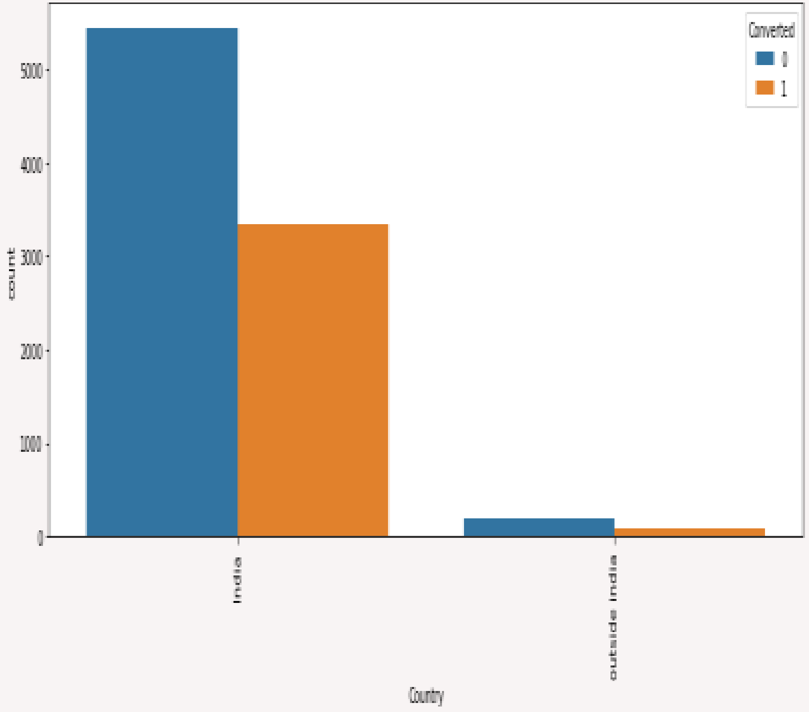
You were given a leads dataset from the past that contained approximately 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on the Website, Total Visits, Last Activity, etc., which may or may not be useful in ultimately deciding whether a lead will be converted or not. The target variable, in this case, is the column ‘Converted’, which tells whether a past lead was converted or not, where 1 means it was converted and 0 means it wasn’t converted. You can learn more about the dataset from the data dictionary provided in the zip folder at the end of the page. Another thing that you also need to check out is the levels present in the categorical variables. Many of the categorical variables have a level called ‘Select’ which needs to be handled because it is as good as a null value (think why?).

Goals Of Case Study

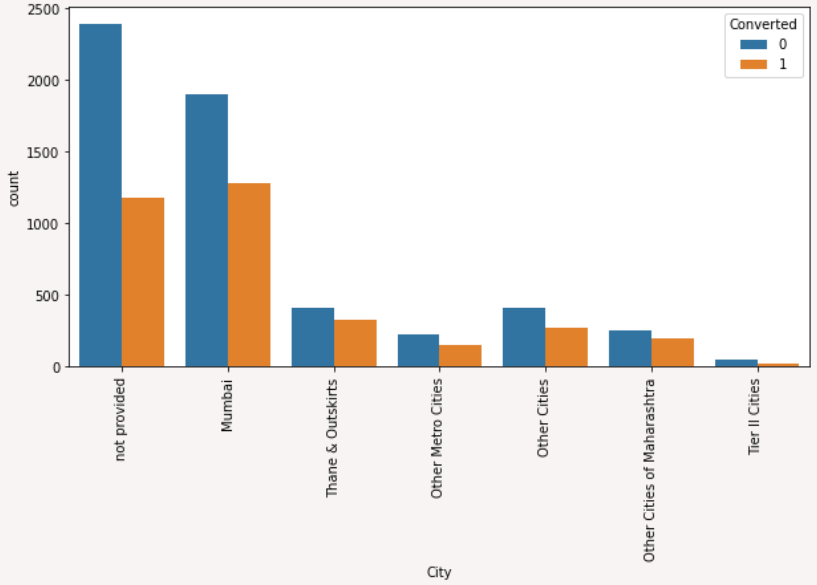
1. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads, which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e., most likely to convert, whereas a lower score would mean that the lead is cold and will mostly not get converted.
2. There are some more problems presented by the company that your model should be able to adjust to if the company's requirements change in the future, so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it out based on the logistic regression model you got in the first step. Also, make sure you include this in your final PowerPoint presentation, where you'll make recommendations

Analysis Done

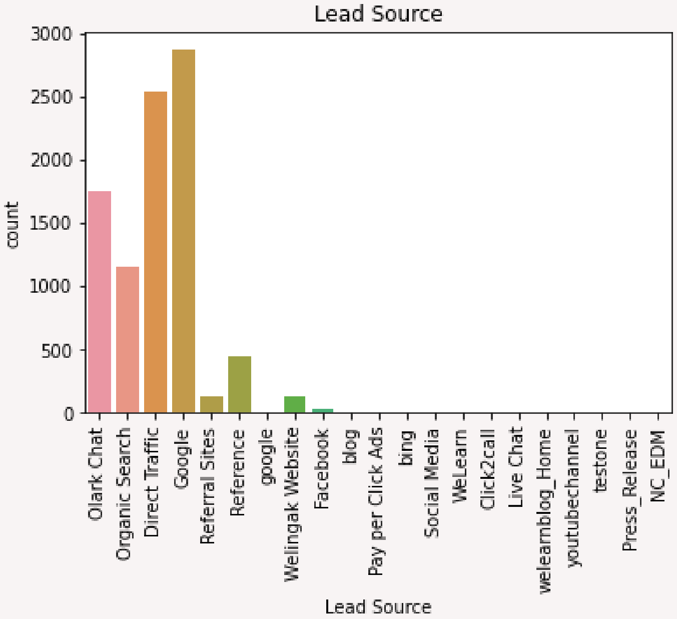
plotting spread of Country columnn to determine the number of foreign users and indan users



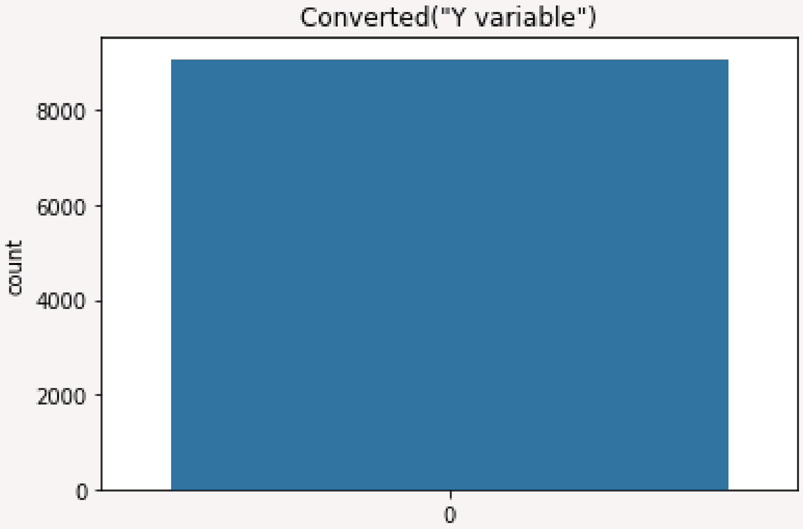
This graph visually represents the distribution of leads across various cities and their conversion status. It is created using Python's Matplotlib and Seaborn libraries, with the figure size set for optimal presentation. The graph, known as a count plot, uses bars to show the frequency of leads in each city, colored based on whether they are converted or not. The x-axis tick labels are rotated for improved readability, especially when dealing with numerous cities. Overall, this graph effectively communicates insights about lead distribution and conversion rates by city, making it a valuable tool for decision-making in marketing and sales strategies.



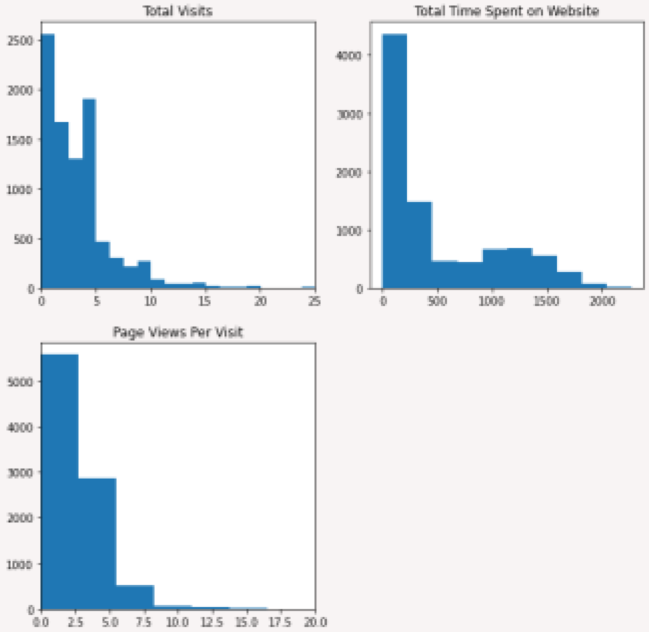
The 'Lead Source' column from the dataset 'Lead_data' is plotted on the x-axis, showing the number of leads attributed to each source. To enhance readability, the x-axis tick labels are rotated by 90 degrees. The plot is titled 'Lead Source' to clearly indicate its purpose, and then it is displayed using plt.show(). This code effectively communicates the distribution of leads across different sources, providing valuable insights into where leads are originating from in the dataset.



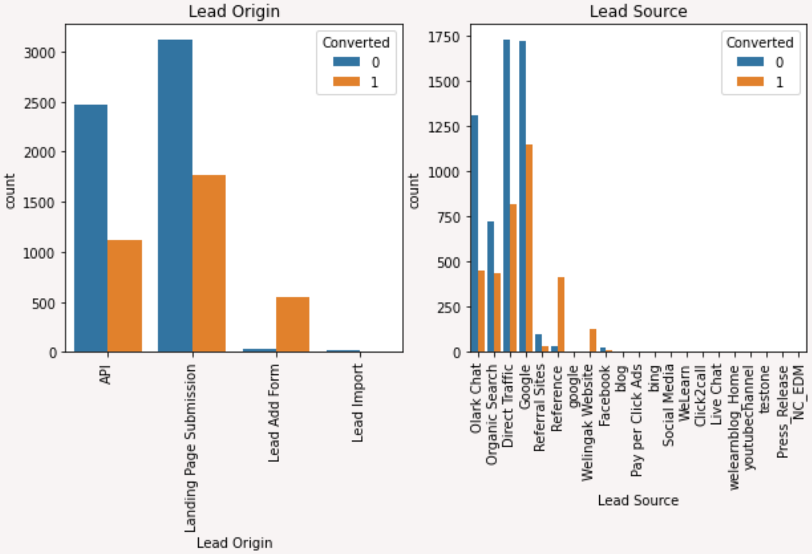
The plot displays the frequency of converted and non-converted leads, with the x-axis representing the conversion status. The plt.title() function sets the plot title as 'Converted("Y variable")', clearly indicating that the variable being analyzed is the dependent variable (Y variable) related to conversion. Finally, plt.show() displays the plot. This visualization provides a straightforward overview of lead conversion status, helping to understand the distribution of converted and non-converted leads in the dataset.



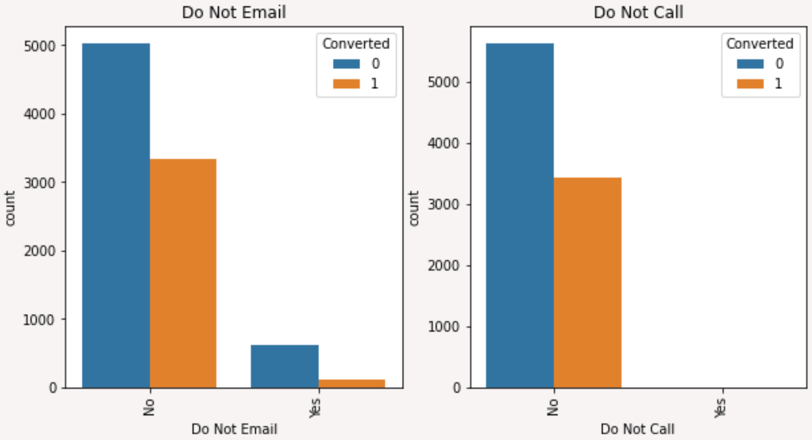
The provided code segment creates a 2x2 grid of histograms using Matplotlib to visualize different numerical variables from the dataset 'Lead_data'. The first subplot displays the distribution of total visits with 200 bins, setting the title as 'Total Visits' and limiting the x-axis to 0 to 25 visits. The second subplot shows the histogram of time spent on the website with 10 bins and is titled 'Total Time Spent on Website'. The third subplot presents the distribution of page views per visit using 20 bins, with the title 'Page Views Per Visit' and an x-axis limit of 0 to 20 page views per visit. These histograms offer a clear overview of lead behavior metrics, aiding in understanding patterns related to total visits, time spent on the website, and page views per visit in the dataset.



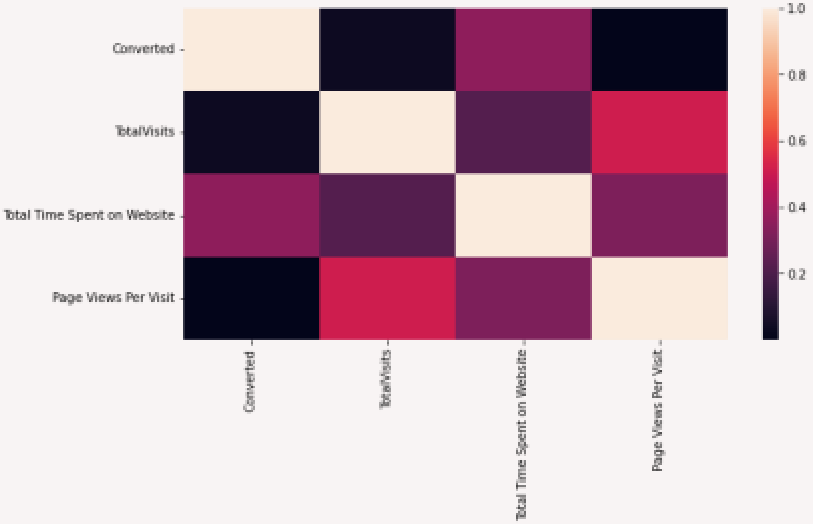
The code generates a 2x2 grid of count plots using Seaborn and Matplotlib, focusing on categorical variables related to lead origin and lead source in the dataset 'Lead_data'. In the first subplot, a count plot of 'Lead Origin' categorized by conversion status ('Converted') is displayed, with x-axis tick labels rotated for better readability. The second subplot showcases a similar count plot for 'Lead Source' categorized by conversion status as well. These plots provide visual insights into the distribution and effectiveness of lead origins and sources in terms of conversion rates, aiding in decision-making for marketing and sales strategies.



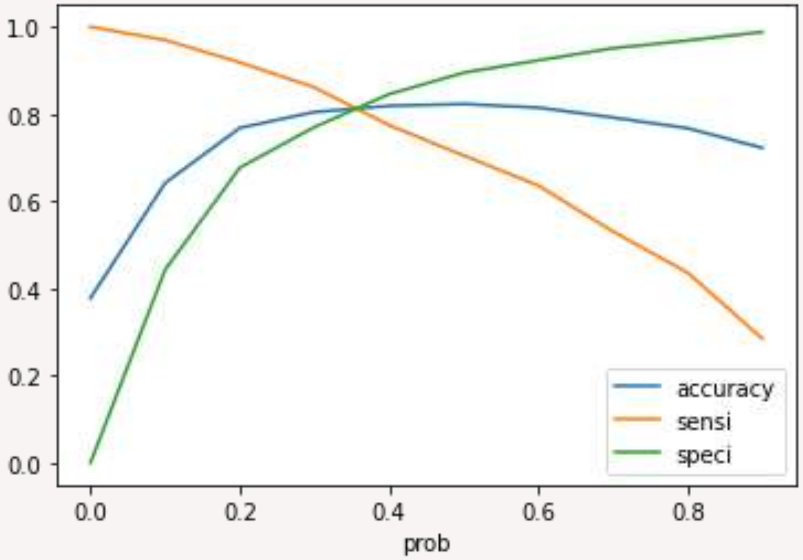
The provided code segment creates a figure with a size of 10x5 and plots two count plots side by side using Seaborn and Matplotlib. The first subplot displays a count plot of the 'Do Not Email' variable categorized by conversion status ('Converted'), with x-axis tick labels rotated for better readability. The second subplot presents a similar count plot for the 'Do Not Call' variable. Both plots visualize the frequency of leads marked as "Do Not Email" or "Do Not Call" and their conversion status, providing insights into the impact of these preferences on lead conversion. The titles 'Do Not Email' and 'Do Not Call' appropriately label each subplot, enhancing the understanding of the plotted data.



The code snippet generates a heatmap using Seaborn and Matplotlib to visualize the correlation among variables in the dataset 'Lead_data'. The heatmap is displayed in a figure with dimensions 10x5, providing a clear overview of the correlation coefficients between different variables. Warmer colors (e.g., red) indicate stronger positive correlations, while cooler colors (e.g., blue) represent negative correlations. This visualization helps in understanding the relationships between variables, identifying patterns, and exploring potential dependencies that can impact lead conversion and other business outcomes.



The provided code segment plots a line chart using Matplotlib to visualize data from the DataFrame 'cutoff_df'. The x-axis represents the 'prob' column, while the y-axis displays three variables: 'accuracy', 'sensi', and 'speci'. The line chart shows the trends or changes in these variables across different probability cutoff values. This type of visualization helps in analyzing the performance metrics (accuracy, sensitivity, specificity) with respect to varying probability thresholds, providing insights into model behavior and performance at different decision thresholds.



Summary

The project for company X Education focused on building a predictive model to increase conversion rates among potential users. Through exploratory data analysis(EDA), columns with high missing values were dropped, NaN values were replaced, and important variables were imputed. The data was split into training and testing sets, scaled, and used for model building using Recursive Feature Elimination(RFE) for feature selection. Model evaluation was conducted using both Sensitivity-Specificity and Precision-Recall metrics, determining optimal cutoff values of 0.35 and 0.44 respectively. The model identified key variables contributing to conversion, providing a reliable prediction tool for the company's decision-making process.