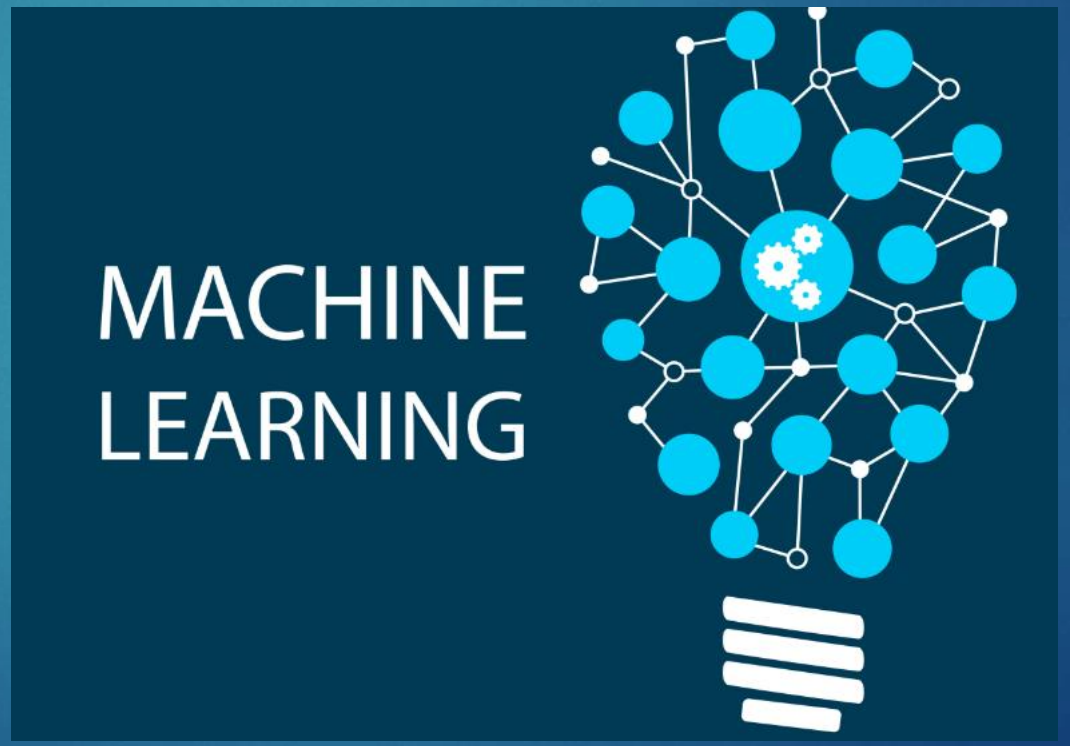
Lead scoring assignment



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 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. 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 (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating 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 of the target lead conversion rate to be around 80%.

Goals for the study :

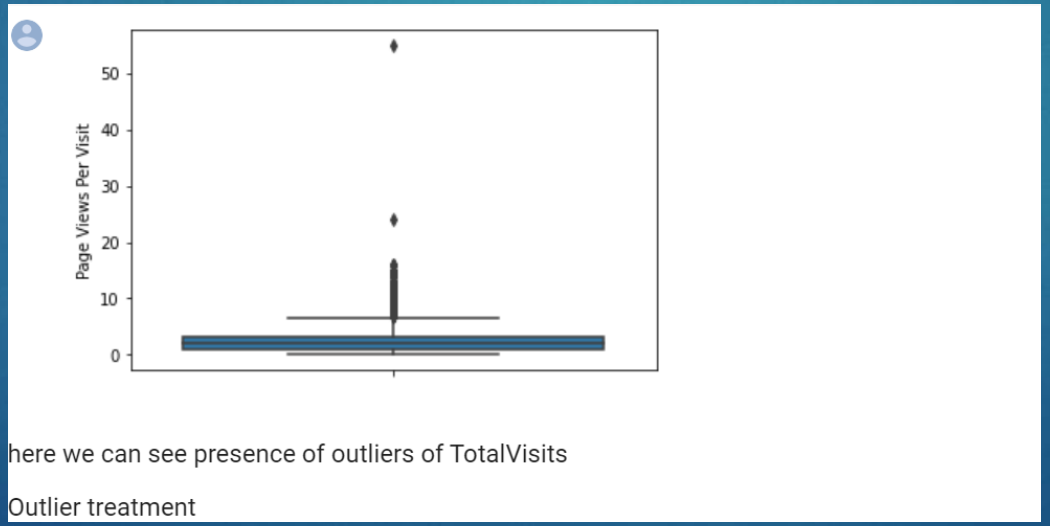
1. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will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

Methods Used for the study :

- Importing python libraries
- Reading and understanding the data
- Data cleaning if needed
- Exploratory data analysis
- Data preparation
- Model Building
- Model Evaluation and conclusion

Reading and understanding the data with Data cleaning :

- First we import python libraries and then check for rows and columns .
- Then we check for null values and duplicate values if there are any . Then check for outliers .
- Then we drop the unwanted values

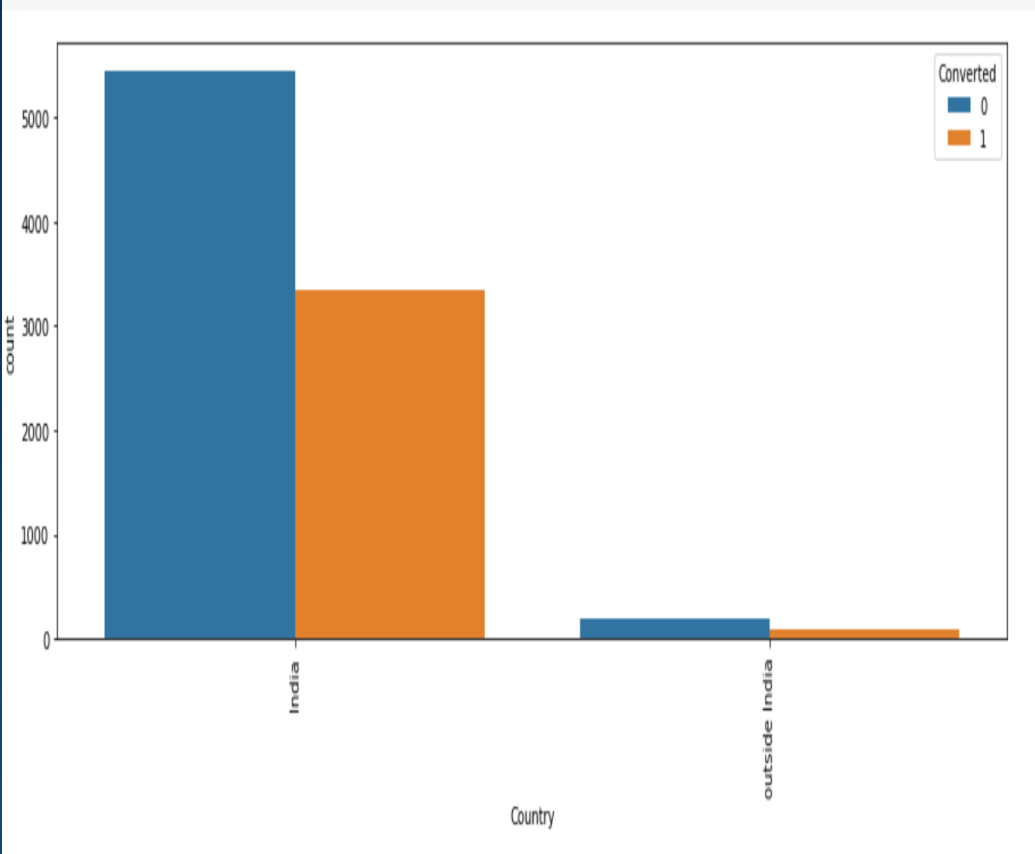


here we can see presence of outliers of TotalVisits

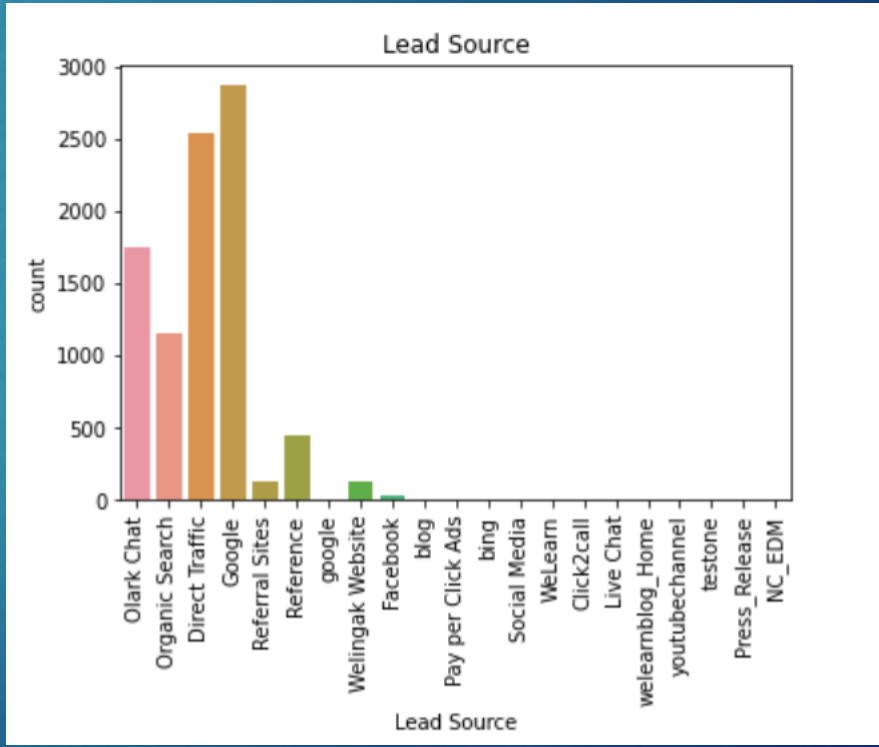
Outlier treatment

Exploratory Data Analysis :

- Data Imbalance
- Graph Functions
- Univariate analysis
- Bivariate analysis

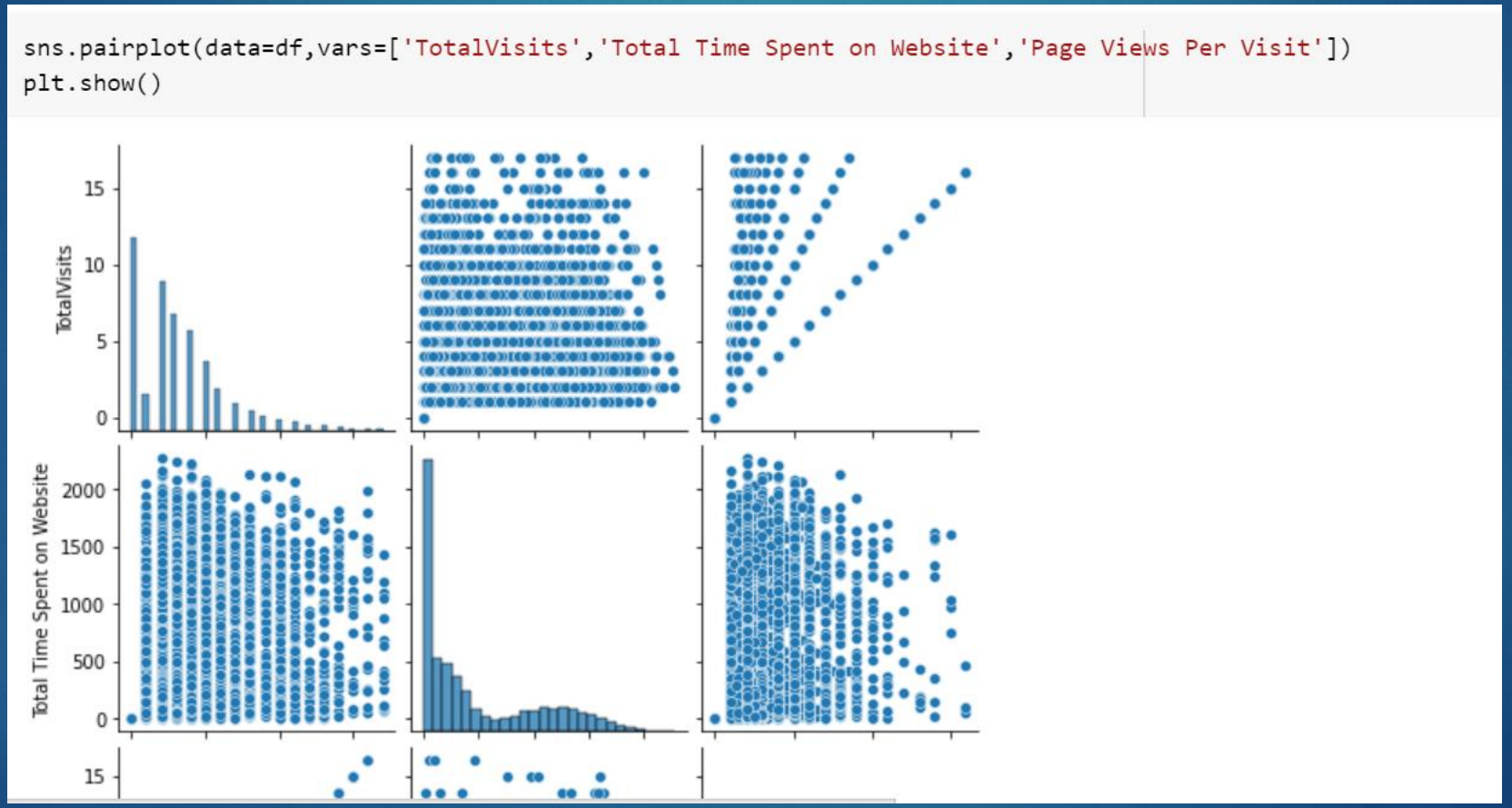


EDA With graphs : We study various charts and perform EDA on the data.

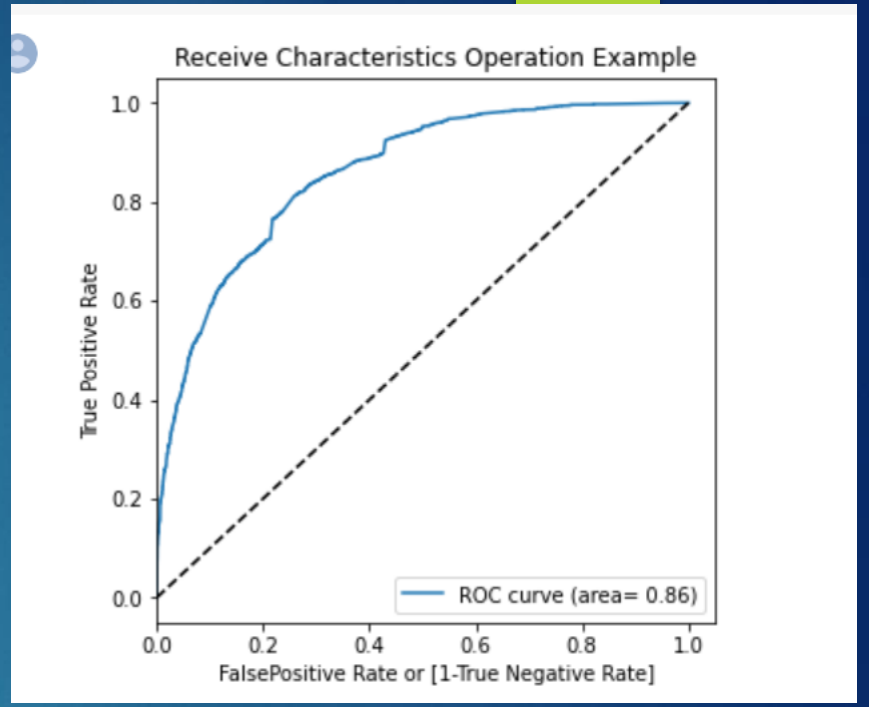


We use sns pair -plot to see the data :

```
sns.pairplot(data=df,vars=['TotalVisits','Total Time Spent on Website','Page Views Per Visit'])
plt.show()
```



Model Building :



- We determine the dummy variable .
- We then train – test split the data .
- Then we start our model building , finding p and vif.
- After thorough analysis , prediction is being made .
- Model Evaluation done with graph.
- Then with different cutoff values , accuracy and precision , recall is being calculated.

Conclusion:

We can use the lead_score column to identify which potential leads to prioritize first. The higher the score, the higher chances are there for the lead to convert. If there are limited sales representatives, then score cut-off should be higher to ensure a higher conversion probability people are contacted further to turn them into a potential customer. It is the same as increasing the precision value of the model by adjusting the cut-off point to a higher value. In case there are more resources available in the sales team (i.e., interns, etc.), then the score cut-off can be lowered. As there are more human resources, the company can afford a higher rate of False positives as it will increase the customer outreach and, in turn, increase the potential customer who will take the online courses.