LEAD_SCORE_CASE_STUDY Summary:

Goals of the Case Study:

- 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.

ANALYSIS APPROACH:

- **Reading and Data Understanding :**
- **Data Cleansing:**
- Checking Null Values/ Treatment
- **Data Visualization:**

  Univariat Analysis

  Multivariat Analysis
- **Data Preparation For Modelling:**

  Scaling Data (Standardization)
- **Model Building:**

  Logistic Regression and Feature selection
- **Final Analysis And Business recommendations:**

Reading and Data Understanding:

- Here we loaded the dataset and saw its shape, Info and descriptive stats of numerical features.

This dataset has:

- 9240 rows,
- 37 columns

- **There are No duplicate values in Lead Number:**
    - Clearly **Prospect ID** & **Lead Number** are two variables that are just indicative of the ID number of the Contacted People and we dropped them.
- From Info and describe function we came to know that there so many Null and outliers are there in data set.

Data Cleansing:

In this particular step we saw Null values for each Features and we treated them accordingly. Null value counts and how we treated them are given below:

- **Lead Source - 36**
  We **replaced Nan Values and combining low frequency values.**

- **TotalVisits - 137**
  Dropped all the rows which have Nan Values. Since the number of Dropped rows is less than 2%, it will not affect the model.

- **Page Views Per Visit - 137**
  Dropped all the rows which have Nan Values. Since the number of Dropped rows is less than 2%, it will not affect the model.

- **Last Activity - 103**
  We **replaced Nan Values and combining low frequency values.**

- **Country - 2461**
  We **Imputed this feature's Null values to India As IT was most occurring country. Then we dropped that feature because majority (97%) of leads are from India.**

- **Specialization - 3380**
  Lead may not have mentioned specialization because it was not in the list or maybe they are a students and don't have a specialization yet. **So we will replace Null values here with "Not Specified"**

- **How did you hear about X Education - 7250**
  We **dropped this column** because it has more than **75% of data is missing.**

- **What is your current occupation - 2690**
  From data dictionry we know it Indicates whether the customer is a student, umemployed or employed so **imputed Nan values with highest occurring category "Unemployed"**

- **What matters most to you in choosing a course - 2790**
  It's an option selected by the customer indicating what is their main motto behind doing this course so we **replaced Nan values with Mode "Better Career Prospects"**

- **Tags - 3353**
  From data dictionary we know that Tag is nothing but current status of Leads so **We imputed Null values and some less value counted category to "Not_specified"**

- **Lead Quality - 4767**
  We **dropped this column** because it has more than **45% of data is missing.**

- **Lead Profile - 6855**
  It's lead level assigned to each customer based on their profile **we dropped it because it is having more than 70% data.**

- **City -3669**
  We **imputed City with Mumbai** because **Mumbai is most occurring city.**

- **Asymmetrique Activity Index -4218**
  We **dropped this particular variable** because of the **percentage of Null values is More (~45%)**

- **Asymmetrique Profile Index -4218**
  We **dropped this particular variable** because of the **percentage of Null values is More (~45%)**

- **Asymmetrique Activity Score - 4218**
  We **dropped this particular variable** because of the **percentage of Null values is More (~45%)**

- **Asymmetrique Profile Score – 4218**
  We **dropped this particular variable** because of the **percentage of Null values is More (~45%)**

Data Visualization:

In this Process we visualized data to understand the features much better and some of the inference are given below:

- Maximum number of leads are From India .
- Number of Values for India are quite high (nearly 97% of the Data), this column can be dropped.
- Majority of Leads are from Mumbai and Their Conversion rate is Low as compared to leads from that city.
- Maximum number of leads are generated by Google and Direct traffic.
  Conversion Rate of reference leads and leads through welingak website is high.
  To improve overall lead conversion rate, focus should be on improving lead converion of olark chat, organic

search, direct traffic, and google leads and generate more leads from reference and welingak website.

- Maximum number of leads are Management_specialisation and One's who not Specified their Work history.

- Conversion Rate of Management_specialisation and One's who not Specified their Work history is high.
- Working Professionals going for the course have high chances of joining it.
- Unemployed leads are the most in terms of Absolute numbers.
- Leads spending more time on the website are more likely to be converted.
- Website should be made more engaging to make leads spend more time.
- API and Landing Page Submission bring higher number of leads as well as conversion.
- Lead Add Form has a very high conversion rate but count of leads are not very high.
- Lead Import and Quick Add Form get very few leads.
- In order to improve overall lead conversion rate, we have to improve lead converion of API and Landing Page Submission origin and generate more leads from Lead Add Form.

From the heat map on the right side we can conclude below points:

- Total Visit And Page Views Per Visit Are Positively Correlated With Correlation Of -0.51
- Total Time Spent on Website And Converted Are Positively Correlated With Correlation Of 0.35
- Total Time Spent on Website And Page Views Per Visit Are Positively Correlated With Correlation Of 0.74

Data Preparation For Modelling:

It in Step we did one hot encoding for categorical variables and Scaled Data (Standardization) for numeric variables. And then Split data for Train and Test set and Separated Predictive and Target Variables.

Model Building:

Here in this step we initialised our base logistic regression model with 56 features then we used **RFE** for **Dimesionality Reduction (Feature elemation)** and reduced features to 15, then we checked **VIF** for **Multicolinearity** and dropped some features which highly multicolinear**,** then checked **P-score** for spignificance of particular features of all features and dropped so redundant features.

Then we came up with these features:-

- **Lead Origin_Lead Add Form**
- **Tags_Will revert after reading the email**
- **Last Activity_SMS Sent**
- **Last Notable Activity_Modified**
- **Lead Source_Direct Traffic**
- **Lead Source_Welingak Website**
- **Tags_Other_Tags**
- **Total Time Spent on Website**
- **Tags_Closed by Horizzon**
- **Tags_Ringing**
- **Tags_Interested in other courses**
- **Tags_Lost to EINS**
- **Last Notable Activity_Olark Chat Conversation**.

Model Evaluation:

For model evaluation we used so many sklearn metrics and those are accuracy score, confusion matrix, ROC_score, Precision & Recall score and we checked sensitivity and specificity.

And for Optimal **Cutoff point Point** we checked both ROC and accuracy sensitivity and specificity for various probabilities and we saw that kept 0.3 as optimum point to take it as a cutoff probability.

Here is our models Observations report for our Train data set:

The ROC curve has a value of 0.97, which is very good. We have the following values for the Train Data:

Accuracy : 92.29%

Sensitivity : 91.70%

Specificity : 92.66%

Precision score: 0.88

Recall Score : 0.92

Confusion matrix:
```
[[3597,  285],
 [ 198, 2187]]
```

Here is our models Observations report for our Test data set:

After running the model on the Test Data these are the figures we obtain:

Accuracy : 92.78%

Sensitivity : 91.98%

Specificity : 93.26%

Precision score: 0.89%

Recall Score : 0.91%

Final Analysis And Business recommendations:

The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model**.**

Business recommendations:

- The company **should make calls** to the leads coming from the `lead sources "Welingak Websites"` and `"Reference"` as these are more likely to get converted.
- The company **should make calls** to the leads who are the `"working professionals"` as they are more likely to get converted.
- The company **should make calls** to the leads who spent `"more time on the websites"` as these are more likely to get converted.
- The company **should make calls** to the leads coming from the `lead sources "Olark Chat"` as these are more likely to get converted.
- The company **should make calls** to the leads whose `last activity` was `SMS Sent` as they are more likely to get converted.
- The company **should not make calls** to the leads whose `last activity` was `"Olark Chat Conversation"` as they are not likely to get converted.
- The company **should not make calls** to the leads whose `lead origin` is `"Landing Page Submission"` as they are not likely to get converted.
- The company **should not make calls** to the leads whose `Specialization` was `"Others"` as they are not likely to get converted.
- The company **should not make calls** to the leads who chose the option of `"Do not Email" as "yes"` as they are not likely to get converted.