**Lead Case Study Report**

Following is the approach I took to build the lead score case study and also the learnings I got from it:

1. I load the data into a data frame and understand the data, including checking dimensions and statistical summary, and finding out the number of nulls in each column.
2. Checking for any duplicated row and then dropping the columns' **Prospect ID'** and **'Lead Number.'**
3. After checking the percentage of nulls in each column, the columns with more than 70% nulls were dropped into them.
4. After this, we do exploratory data analysis for each column and either drop the null columns or replace the nulls with the mode of categorical variables.
5. After replacing the null values, check for the rare levels in categorical variables, merge the levels, and visualize the target variable's levels.
6. Also, some of the variables are dropped as they almost have identical values in all the rows. So, they don't make any significance for predicting as well as for any inference, so we should drop such variables. The variables include **Search, Magazine, Newspaper Article, Newspaper, X Education Forums, Digital Advertisement, Through Recommendations, 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**. Other than these columns, we also need to drop **Asymmetrique Activity Index, Asymmetrique Activity Score, Asymmetrique Profile Index, Asymmetrique Profile Score** as they have high variation and 45% nulls value therefore it would be better to drop these columns.
7. Other than categorical variables, we check outliers in numerical variables, do outlier treatment, and visualize the box plot.
8. After visualizing, we have gathered inference about all the categorical variables and numerical variables, which is as follows:
   1. Lead Source
      1. The highest number of people who are converted from Google
      2. Converted to NOT converted ratio is highest for Reference, which means that there are high chances of converting when the leads are sourced from reference
   2. Last Activity
      1. SMS sent has the highest conversion rate, and most numbers of the people got converted who were sent SMS
   3. Specialization
      1. Not specified customers have the highest not opted numbers
      2. All other levels have almost the same conversion rate
   4. What is your current occupation
      1. Most of the count consists of unemployed people, but Working professionals have the highest conversion rate.
   5. Tags
      1. Will revert after reading the email to have the highest count in the column
      2. Closed by horizon have the highest conversion rate while interested in other courses have the lowest conversion rate
   6. Lead Quality
      1. High in Relevance have a high conversion rate
      2. While people who are Not sure have the lowest conversion rate
   7. City
      1. Mostly all the values are Mumbai in the City column, so we can conclude that the service is primarily used in Mumbai
   8. Lead Origin
      1. Lead Add Form has the highest conversion rate, but the count is less
      2. Therefore organizations should focus more on this
   9. TotalVisits
      1. People having higher visits are likely to convert more than lower visits
   10. Total Time Spent on Website
       1. People spending more time on the website have higher chances of conversion
   11. Page Views Per Visit
       1. page views per visit don't have much effect on the target variable as the median is the same for both the categories
   12. A free copy of Mastering The Interview
       1. Even though the No is high, the conversion rate for both Yes and No are the same
   13. Last Notable Activity
       1. The SMS sent to people has the highest conversion rate for the Last notable activity
9. After this, we further create dummy variables for categorical variables and remove the original variables. Also, we map the binary variables with 1 and 0 and then continue for model building with 52 columns.
10. So, first, we rescale the numerical variables using Standard Scaler to converge the model faster.
11. Now let's build the model using RFE from scikit-learn to select the top 15 variables from the data set, remove the variables having insignificant p-values and proceed further.
12. So we removed two such variables, which are **Tags\_Ringing** and **Tags\_switched off,** and predicted the probability for each record
13. Next, we predict the Converted by deciding the probability cutoff (decision boundary) as **0.5**, but that is not the correct way to decide, so we draw the **ROC curve, accuracy, sensitivity and specificity curve, and precision and recall curve** to decide the decision boundary. So, we arrive at a conclusion of **0.3** for the decision boundary.
14. Based on this, we allocate the Lead score to each record in train data and check the VIF values for features, but all the values come out to be less than 5, so we are good here.
15. Next, we do the same steps for test data and predict the probability for test data and by using the decision boundary as **0.3**, decide the predicted Converted and calculate the accuracy, sensitivity, and specificity
16. As we can see, the accuracy for test and train data is approximately the same ≈ 91%, so we can conclude that the model is performing well and not doing overfitting.
17. We can also see from the model that Lead Quality, Tags and Last Notable Activity have more significance towards target variable, so the organization should focus more on such areas.