Lead Scoring Case Study Summary

First effective way of working on leads is to start with the **hot leads** i.e. leads that have higher probability of getting converted. This will not only result in higher conversion ratio but also effective use of time. Time spent on nurturing hot leads can be increased but whereas time spent on leads with low score (cold leads) can be minimized.

Determining hot and cold leads can be done by using a **logistic regression model**. Using the meta data provided for each lead, we will build a logistic regression model and assign **lead score** to each lead.

Current conversion rate of the leads is **39%**.

EDA is extremely important step prior to building model. Some of the Key insights from EDA helps in treating the data correctly. Data cleaning helps in building efficient model. Steps like missing value imputation, scaling, outlier treatment must be performed at minimum to ensure quality of data is not compromised.

By using Boxplot and descriptive statistics we can indicate that there are outliers in the dataset.Removing the outliers will result approximately in **9% data loss**.

Bivariate Analysis

- Lateral students and the visitors showing interest on next batch have higher chances of getting converted.
- Lead quality tagged with “High in Relevance” has high conversion rate history.
- Lead originated through “Lead Add Form” and “Quick Add Form” has high possibility of getting converted.
- Lead belongs to Selinga Website, WeLearn, Live Chat and NC_EDM converts more than any other sources.

Plotting a heatmap indicates some correlation between variables. VIF will be further used during model building – we see high correlation among below attributes.

- **Categorical column**

Search, Newspaper Article, X Education, Digital Advertisement, Through Recommendations

- **Numerical Column**

Total Visits, Total Time Spent on Website, Page Views Per Visit.

Final Model and Interpretation

Model6 is a good model based on statistical summary.

- Final model contains 12 most important features which satisfy all the selection criteria.
- Lead score having conversion probability greater than 0.43 are being predicted as “Converted”.
- Using this probability threshold value (0.43), the leads from the test dataset have been predicted whether they would get converted or not.

- Confusion matrix with cut-off 0.43, to calculate evaluation metrics.

Evaluation Metrics

- Since the ROC curve is close to the upper left part of the graph, it means this model is a very good model (The value of AUC for our model is 0.93)
- Tradeoff between sensitivity and accuracy can be observed (cutoff = 0.34).
- Ideal cutoff of 0.43 is observed from recall and precision plot.

Conclusion and Recommendations:

Followings are top three features that contribute to decision which mean the conversion probability of a lead increases with increase in values of these features:

- Lead Origin
- What is your current occupation
- Last Activity

Top three categories that contribute to decision.

- Lead Origin -> Lead Add Form
- What is your current occupation -> Working Professional
- Last Activity -> SMS Sent

This model will help to identify the hot leads which would enhance speed to lead and the response rate. Approaching only to hot lead would result in

- Shorter sales cycle through intuitive prioritization
- Control over volatile buying cycle
- Better opportunity-to-deal ratio
- Minimize opportunities loss
- Increase in revenue
- Increase marketing effectiveness
- Better sales forecasting