LEAD SCORING CASE STUDY SUMMARY

SUBMITTED BY: Vivek Tiwari (DSC64 BATCH)

Problem Statement

1. X Education company sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%

2. 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. Achieve lead conversion rate to be around more than 80%.

3. 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.

4. 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.

5. Finding the Top three variables in your Model which contribute most towards the Probability of a Lead Getting Converted.

The following technical steps are used for the analysis.

1. **Data cleaning, Manipulation (Missing value treatment) & outlier analysis**
2. **EDA**
3. **Data Preparation for Modelling**
4. **Model Prediction**
5. **Model Evaluation**
6. **Lead score calculation**
7. **Hot Leads Determination**
8. **Feature Importance Determination**
9. **Business Insights**
10. **Results & Recommendation**

CONCLUSION & RECOMMENDATIONS

After trying several models, we finally chose a model no 5 with the following characteristics:

- All variables have p-value < 0.05, showing significant features contributing towards Lead Conversion.
- All the features have very low VIF values, means hardly there is any multicollinearity among the features. This can be seen from the heat map.
- The ROC curve has a value of 1, which is very good!
- The overall accuracy of Around 80% at a probability threshold of 0.34 on the test dataset is also very acceptable.

- **For Train Dataset**
  - Accuracy : 80.33%
  - Sensitivity/Recall : 81.66%
  - Specificity : 79.50%
  - False positive rate - predicting the lead conversion when the lead does not convert: 0.20
  - Precision/Positive predictive value: 71.08%
  - Negative predictive value: 87.54%
  - ROC : 1
  - F1 Score : 0.76

- **For Test Dataset**
  - Accuracy : 80.36%
  - Sensitivity/Recall : 81.70%
  - Specificity : 79.48%
  - False postive rate - predicting the lead conversion when the lead does not convert: 0.20
  - Precision/Positive predictive value: 72.10%
  - Negative predictive value: 87.00%
  - ROC : 1.0

- **The optimal threshold for the model is 0.34** which is calculated based on trade-off between sensitivity, specificity and accuracy. According to business needs, this threshold can be changed to increase or decrease a specific metric.
- High sensitivity ensures that most of the leads who are likely to convert are correctly predicted, while high specificity ensures that most of the leads who are not likely to convert are correctly predicted.