SUMMARY

LEAD SCORING CASE STUDY

This analysis is done for X Education and to find ways to get more industry professionals to join their courses. The basic data provided gave us a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate.

The following are the steps used:

1) Data Information:

- The data set that we are working on contains 37 features and 9240 rows.
- From the data set our dependent feature was 'Converted' on which we have to build model using the rest 36 features.

2) Data Pre-Processing:

- Null Values: - We have removed those features which are having more than 50% of null values. We have removed those rows which contains less than 2% of null values in that features.
- For null values in categorical features we have replaced null values with the mode (maximum occurrence) value.
- Also we dropped those features which was insignificant or unnessary for model building.

3) Exploratory Data Analysis:

- Univarient Analysis:

  - In the feature 'Converted' the ratio of customer converted to that of not-converted was 37%.
  - Majority of the lead was originated through 'Landing Page Submission'.
  - Major sources of lead were 'Google' and 'Direct traffic'.
  - Majority of the customers were unemployed in 'Current Occupation' feature.
  - Better Carrier Prospects was the main reason behind choosing this course.
  - More than 90% of the customers were from India.

- ❖ Bi-lateral Analysis:
  - As the time spend on website increases more than 500, the probability of the customers converted increases.
  - Those customers who come through ‘Reference’ and ‘Welingak Website’ from lead source have very high chances of converting.
  - Majority of the customers whose last activity was ‘SMS Sent’ had higher chances of converting.

4) Model Building:

- ➢ First we have divided the dataset in to dependent and independent features.
- ➢ Then we spitted the dataset into train and test data in the ratio of 70:30.
- ➢ With the help of standardization in all the continuous features we brought we brought their values into same scale.
- ➢ Then with the help of Recursive Feature Elimination we have selected the top 15 features which will predict our dependent variable.
- ➢ Than from those 15 features we further reduce to those features whose P-value and Variation Inflation Factor was less than 0.05 and 5.
- ➢ Then with those features we have predicted the dependent variable.

5) Model Evaluation:

- ➢ With the help of R.O.C curve we have selected the threshold point for predicting the provability of customer who can convert.
- ➢ We then predicted our model on the data and compared it with the original converted data and found the following result

  Train dataset:
  - Accuracy = 79.36%
  - Sensitivity = 80.64%
  - Specificity = 78.39%
- ➢ Same process we have done on test data and got the following result

  Test dataset:
  - Accuracy = 78.32%
  - Sensitivity = 78.87%
  - Specificity = 77.32%
- ➢ Finally we have created a dataset which can give a lead_score (Probability value *100) for predicting the probabilities of customers to be converted.