LEAD SCORE CASE STUDY SUMMARY REPORT

Business understanding:

X Education company sells online courses for professionals, later those interested will check the sites and fill the form if they want further details or enroll in it. This company checks the individual's activity and even through past referrals, find the prospect leads. Then the sales teams, through different modes contact them and convert the lead into paying customers.

Problem statement:

Presently the conversion rate is 30% and the company wants us to build a model which can help them increase this rate.

Solution approach:

We have chosen to 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.

Summary report for the business:

The business can calculate the Lead score of each candidate using the optimal cut off and if the lead score is greater than 35 (as the cut off probability is 0.35) then that candidate can be considered as a Hot Lead and the sales team can concentrate more on that person to convert him into a potential paying customer.

To increase the lead score of the candidate the business can target the Lead source feature and try to increase the count of categories Welingak Website and Reference as it has more potential leads. Moreover, the business can target on working professionals as they can turn into possible prospect lead. The business needs to target less on leads who have selected the 'Do not email' option as there is a negative coefficient and this indicates that it doesn't help the business.

If there is business decides to the increase the budget for sales team then it can reduce the probability cut off from 0.35 i.e., can target customers with lead score less than 35 as this change would lead to an increase in the targeted customers, and it might result in high lead conversion rate in turn improving the profits of the business.

And if the business wants to concentrate on any other vertical and intends to reduce the sales team work (thereby diverting the employees towards a new vertical, if any) then it can choose to increase the cut off from 0.35 to a higher one. This strategy can be applied when the business has reached its target and is willing to accept less lead conversion rate. During the application of this strategy, make sure that you target at least those with lead source of welingak and reference categories and working professionals as these have very high prospects of getting hot leads and potential paying customers.

To conclude the conversion rate of the leads is 80% when the lead with lead score 38 and above is targeted.

Detailed Steps taken to solve this problem (The data analyst team can get better insights from this):

1. Firstly, we have imported the data and the libraries needed to build this model
2. Data Cleaning:
   2.1 Firstly, converted the columns have "Select" as their values to null as this gives us the information that either the user hasn't selected any option or chose not to select.
   2.2 Have dropped the columns with greater than 45% missing values and even with one unique value
   2.3 Performed binary mapping
   2.4 Used the appropriate data imputation technique for other columns with missing values.
3. Data Analysis:
   3.1 Identified the outlier data and handled it.
   3.2 Performed the EDA on all the columns and have plotted them against 'Converted' and inferred which columns might be suitable for further analysis
4. Data Preparation:
   4.1 Have created a dummy variable for the categorical columns
   4.2 Performed the test train split
   4.3 Feature scaling have been done
5. Have started building the logistic regression model and using RFE technique chose the top 20 variables which will give more accurate results.
6. Dropping the columns p values greater than 0.05 and VIF greater than 2.5
7. Calculated the model evaluation parameters like accuracy, sensitivity, specificity, Precision and recall
8. The optimal cut off was 0.35 so we have considered it as a cut off probability
9. Make predictions on test data and using the same threshold value of 0.35 predicted if the lead is converted or not.
10. Adding of a lead score column to the final model which is populated by multiplying the conversion probability value with 100 so that we get the score between 0-100.