**Summary: Lead Scoring Case Study**

We carried out this analysis for the X education company to find out potential leads that can convert and improve their conversion rates. The basic data that the company provided had information like customer visits to a site, time they spend there, how they reach the site and conversions.

With the help of the below steps, we solved this problem for the X education company.

**Data cleaning**

There were columns with a majority of null values and columns with few null values. For columns with majority of null values, we removed those columns as they may disturb the model. For columns with few null values, we just removed the rows where the null values were found and then carried on with the exploratory data analysis. We removed the entire columns that had more than 35% of null values. Also, we checked if there were duplicate values and unique value counts. We removed columns with one unique value.

**Exploratory data analysis**

We carried out the EDA for this dataset and checked how categorical variables and numerical variables were related to conversions using basic visualizations. Wherever there were outliers, we removed them.

**Dummy variables treatment**

We created dummy variables for the different columns that were categorical. Wherever we found columns not provided, we removed them.

**Train-test split**

We split the pre-processed data into training set and test set. We assigned 70% of data as training set and 30% as test set randomly. Also, we scaled the numerical variables with min-max scaling method for making the model easier to analyse and predict.

**Model building**

We used Recursive feature elimination, an automated feature elimination method to filter out the top 15 features from more than 60 features. Later, we removed the remaining variables with the help of p-values generated using logistics regression and Variance inflation factor. (p-value <0.05) and VIF<5.

**Model evaluation**

We tried to find the optimal cut-off point that we can use to predict if the lead is converted or not. We found this optimal cut-off point with the help of ROC curve. We then tested it with accuracy, sensitivity and specificity. It was around 80 percent.

**Prediction**

We finalized the optimal cut-off point as 0.42 for test and training data and achieved around 80 percent accuracy, sensitivity and specificity.

**Precision-Recall trade-off**

We used the precision-recall trade-off and got a new optimum cut-off point as 0.44. We then found the precision and recall around 75 percent for both test and training data.

**Findings from the model building and evaluation**

Here are the crucial variables that were found to contribute to leads converting for your education course products.

* The total time spent on website
* The total number of visits
* Lead sources like google, direct traffic, organic search and Welingkar website
* Last activity from SMS and Olark chats
* Lead origin is from Lead add format

Focusing and improving on these variables, X education can improve their lead conversion rates and improve their sales.