Lead Scoring Case Study

The problem statement involves an education company, X Education, that aims to improve their lead conversion rate by identifying the most potential leads, also known as 'hot leads'. The company wants to develop a lead scoring model that can accurately predict the conversion chance of each lead based on historical data and various features of the leads, such as their source and profile. The end goal is to provide the sales team with a prioritized list of leads that are most likely to convert into paying customers, enabling them to focus their efforts on these leads and improve the lead conversion rate.

The following was the approach taken:

1. Understanding and loading data
2. Data cleaning - identifying and correcting or removing errors, inconsistencies, and discrepancies in data to improve its quality and prepare it for analysis.
3. Exploratory Data Analysis (EDA) - analyzing and summarizing the main characteristics of a dataset to better understand its underlying structure and patterns.
4. Data Preprocessing - cleaning, transforming, encoding, and scaling data.
5. Splitting the data 70:30 into test and train datasets
6. Model Building using RFE and statsmodel
7. Model Evaluation using VIF, Precision and Recall
8. Making predictions on the test dataset

These are the steps we followed for our assignments:

1. Data Cleaning: Removed redundant features, changed 'Select' labels to null, removed columns with over 45% null values initially, imputed missing values, and fixed label format issues.
2. Data Transformation: Converted multicategory labels to dummy variables, binary variables to '0' and '1', checked and binned outliers, and removed redundant and repeated columns.
3. Data Preparation: Split dataset into train and test, scaled data, and dropped correlated variables.
4. Model Building: Created model using RFE using 15 variables, evaluated model scores found optimal probability cutoff, checked precision and recall, made predictions, and evaluated model on test set.
5. Conclusion: Model has acceptable accuracy and recall/sensitivity, is adaptable to future changes, and top features for good conversion rate are ‘Grouped Tags_EINS / Others’, ‘Last Notable Activity_SMS Sent’ and ‘Last Notable Activity_Unsubscribed’.

Summarized learnings:

1. References and Welingak Website have the highest conversion ratio – leads need to be increased
2. Olark Chat and Direct Traffic have a high number of leads but poor conversion – follow-up is required in this area as there is a potential for conversion
3. The average time spent on the website is ~8 minutes which is slightly shy of the 10 minute mark beyond which leads have more than a 50% probability of being converted; perhaps the website needs to be a bit more engaging and containing more information
4. SMS' work better than other forms of communication / involvement - should look to target customers using this means
5. Target working professionals as they seem to be opting for it more than the others
6. Individuals who were not reachable were quite a big chunk so it might help if they are approached via other means viz. SMS, email, etc. or an alternate contact number is sought