Summary Report: EDA and building Logistic Regression model for Lead Scoring

1. Exploratory Data Analysis (EDA)

1.1 Dataset Overview

The dataset contains a total of 37 rows and 9,240 columns.

1.2 Data Cleaning

The following data cleaning steps were performed:

- Duplicate data was checked for and handled.
- NA values and missing values were checked and handled.
- Columns with a large amount of missing values and deemed not useful for the analysis were dropped.
- Imputation of values was performed where necessary.
- Outliers in the data were checked and handled.

1.3 Univariate Analysis

Univariate analysis was conducted to understand the individual variable distributions and characteristics.

1.4 Bivariate Analysis

Bivariate analysis was performed to examine relationships and correlations between variables.

1.5 Multivariate Analysis

Multivariate analysis was carried out to explore the interactions and patterns between multiple variables.

2. Logistic Regression Model Analysis

2.1 Data Preparation

- Single value features like "Magazine," "Receive More Updates About Our Courses," "Update me on Supply Chain Content," "Get updates on DM Content," "I agree to pay the amount through cheque," etc., were dropped.
- The "Prospect ID" and "Lead Number" columns, which were not necessary for the analysis, were removed.
- Features with low variance, such as "Do Not Call," "What matters most to you in choosing a course," "Search," "Newspaper Article," "X Education Forums," "Newspaper," "Digital Advertisement," etc., were dropped.
- Columns with more than 45% missing values were dropped.

2.2 Model Building and Evaluation

The data was split into training and testing sets using a 75:25 ratio. Recursive Feature Elimination (RFE) was performed to select the top 15 variables. The model was built by removing variables with a p-value greater than 0.05 and a VIF value greater than 5. Predictions were made on the test dataset, and the overall accuracy of the model was determined to be 92%.

2.3 Model Interpretation

Based on the logistic regression coefficients, the key features that significantly influence the prediction of the target variable (y) are as follows:

1. Total Time Spent on Website: For every unit increase in the total time spent on the website, the odds of the target variable (y) being positive (converted lead) increase by approximately 3-fold.
2. Lead Origin_lead add form: Leads generated through the lead add form have a 3.9-fold higher odds of being converted compared to other lead origins.
3. Lead Source_olark chat: Leads generated through the olark chat source have a 2.5-fold higher odds of being converted compared to other lead sources.
4. Lead Source_welingak website: Leads generated through the welingak website have a 75-fold higher odds of being converted compared to other lead sources.

2.4 Recommendations

Based on the EDA and logistic regression analysis, we can provide following recommendations:

1.Should focus on sending attractive offers to individuals who already spend a lot of time on website.

2.Should email individuals who are unemployed about job guarantee or interview guarantee programs

3.Should optimize search engine in google, so when people search for certifications or educations X education shows at the top search.

4. Should also send informative content like free courses and master classes through email

5.Should also start a campaign to encouraging women who are housewives to get educated through courses and provide interview guarantee programs.