Lead Scoring Case Study Summary

An education company named **X Education** sells **online courses** to **industry professionals**. The typical **lead conversion rate** at X Education is around **30%.** The company thinks that the conversion rate is **very low** and wants to **improve it**. To do so, the company wishes to identify the most potential leads called **'Hot lead'**. To identify these hot leads, we need to **build a model** to find features **contributing more** on the **lead conversion rate**. Here, we need to find whether the lead will **take the course** or not. Therefore, the **machine learning algorithm** we have to use is **logistic regression** because it follows a **classification technique**. We're using a **lead conversion past dataset** consisting of **9000 data points**. The **target variable** used for building the model is **'Converted'**.

The first step of any **model building** is to **understand the data**. There are **9240 data points** and **37 features** in the original dataset. The next important step is to **clean the data** for **missing values** and prepare the data for **model building**. The features having **more than 30% of missing values** are **dropped**. Another thing to consider is some features have the category **'Select'** which occupies the **majority of data**. We need to consider this as a **missing value**, and the elegant way to handle it is by **dropping it** while creating **dummy-variables**. We also need to **drop features** having **single majority** (more than **95%**) categories. For example, features like **'Magazine', 'Newspaper', 'Search'**, etc., have single majority category i.e. **'No'** (nearly **99%**). For **categorical variables**, **impute missing values** with **mode value** and for **numerical variables**, impute with **median value**. Make sure that there is **no missing value** in the dataset.

The third step is **Exploratory Data Analysis (EDA)**. From **univariate analysis**, it is found that more **unemployed people** are planning to take the course, and the **lead source** is through **Google**. It is also found from **box plots** that there are **outliers** present in numerical variables like **'TotalVisits'** and **'Page Views Per Visit'**. From **bivariate analysis**, it is understandable that the **top 3 lead sources** for converting the leads are through **Live Chat**, **WeLearn platform**, and **Welingak Website**. The more **converted leads** are identified by **Quick Add Forms** and **Lead Add Forms**.

More **Healthcare Management** workers are converted to **hot leads**. Finally, from **multivariate analysis**, it is evident that the one who **spent more time on the website** were **highly converted** or taken the course because there is **high correlation** between features **'Total Time Spent on Website'** and **'Converted'**.

The next step is to create **dummy variables** for **categorical features**. The categorical features in this case are **nominal** in nature, therefore, used **“One-Hot Encoding”** technique to create **dummies**. The dataset is **split** for **train-test data validation** with a size of **70% training data** and **30% test data**. The **numerical variables** are scaled using **“MinMaxScaler”** scaling technique. Now, the data is ready for **model building**. Use **recursive feature elimination**, select **top 15 features**. Build the model repeatedly by eliminating features with **p-value greater than 0.05** and check **multi-collinearity** using **variance inflation factor** greater than 5. Totally, **four models** were built. At last, **12 columns** were left for **model evaluation**.

The final step is to **evaluate the model** with two methods. The first method consists of **accuracy**, **sensitivity**, and **specificity**. The second method consists of **precision** and **recall**. The area covered by the **ROC curve** is **0.86**, which is perfectly good to continue. The **optimal predicted cutoff** is calculated by **plotting a graph** with different probabilities. The **accuracy** of **training** and **testing data** are **0.790** and **0.797** respectively. The **top 3 categorical/dummy variables** in the model which should be focused on the most to increase the probability of **lead conversion** are:

- **'Lead Source_Welingak Website'**
- **'Lead Origin_Lead Add Form'**
- **'Last Notable Activity_Unreachable'**