LEAD SCORING CASE STUDY - SUMMARY

Problem Statement

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone. X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

Steps Involved in the Case Study

Data Reading and Understanding:

- Number of rows and columns
- Data types of each columns
- Checking first few rows how data looks
- Checking how the data is spread.
- Checking for duplicates, if any

Data Cleaning:

- Checking for any column names correction
- Checking for null values and imputing them with appropriate methods
- We used mode imputation for categorical columns.
- We used mean imputation for numerical columns, if there is no skewness in data.
- We used median imputation for numerical columns, if there is skewness in the data.

Data Visualization and Outliers Treatment:

- We performed univariate analysis on categorical column to see which columns makes more sense and removed those columns whose variance is nearly zero.
- We performed bivariate analysis on categorical columns to see how they vary w.r.t Converted column.
- We performed univariate analysis on numerical columns by plotting box plots to see are there any outliers in the data or not.
- We performed bivariate analysis on numerical columns with Converted column to see how the leads are related to these columns.
- We have used IQR method to treat the outliers in the data set.
- In this step we also plotted the correlation matrix to identify the columns which are correlated.

Feature Scaling

- Columns which have only two levels “Yes” and “No” were converted to numerical using binary mapping.
- Columns which have more than two levels were converted to dummies using pd.get_dummies function.
- Now, the data contained only numerical columns and dummy variables. Before proceeding for model building, we have rescaled all numerical columns by using standard Scaler method.

5. Model Building

We have used Recursive Feature Elimination Technique to remove attributes and built a model on those attributes that remain. RFE uses the model accuracy to identify which attributes (and combination of attributes) contribute the most to predicting the target attribute.

Model Evaluation on Train Set -

we took 0.5 as the cut-of. To confirm that it was the best cut, we calculated the probabilities with different cut-offs.

Final Observations

Train Data:

- Accuracy : 92.29%
- Sensitivity : 91.70%
- Specificity : 92.66%

Test Data:

- Accuracy : 92.78%
- Sensitivity : 91.98%
- Specificity : 93.26%