Lead Scoring Case Study Summary

Problem description

An education company named X Education which provides online courses to industry professionals wants to increase their lead conversion rate.

The company currently 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 current Lead conversion rate is around 30% and the CEO has given a target of 80% lead conversion rate.

Approach:

Below mentioned are the steps used to build the Logistic Regression Model-

1. Data Cleaning and Null value Imputation-
    - Data frame info and description of data was looked at, thereby getting a brief idea of the number of rows and columns, data types of each column and checking how the data is spread.
    - The next step was checking for null values and imputing them. Mode was used imputation for categorical columns and some columns with higher Null value percentage which couldn't be imputed were dropped.
    - Columns which had large number of segments under them were clubbed together, if the values were insignificant as compared to other segments value counts.

2. Exploratory Data Analysis (univariate analysis, outlier detection, checking data imbalance)-
    - Performed univariate analysis on categorical column and derived inferences.
    - Performed univariate analysis on numerical columns by plotting box plots and some variables with outliers were capped.
    - Performed bivariate analysis on numerical columns with Converted column to see how the leads are related to these columns.
    - Correlation matrix was used to identify if the columns were correlated.

3. Data Preparation and Feature Scaling:
    - Dummy variables were created for categorical columns
    - Train-Test split was done in ratio 70:30
    - Before proceeding for model building, Feature scaling numerical columns was done using standard scaler

4. Model Building
    - Using RFE the feature selection was done and top 15 variables were selected, from the 15 again columns were removed based on High P value and High VIF and the optimal model was derived with P. Value less than 0.05 and VIF less than 5.
    - After deriving a stable model, prediction was done on the train data set to get the conversion probability and a new column was created and assigned 1 if probability is greater than 0.5 else 0.
    - After the predicted column was created, using that confusion matrix was derived. Metrics like sensitivity, specificity, precision, recall and accuracy were calculated and ROC curve was plotted to find the area under the curve.

5. Model Evaluation
    - Accuracy, Sensitivity and Specificity was determined for probabilities ranging from 0.1 to 0.9
    - Optimum cut off point of 0.3 was determined from the above table and graph plotted with the same data
    - Predictions were made on the test dataset with cut off .30
    - Below are the observations:

Train Data Metrics

Accuracy	88.87%
Sensitivity	89.36%
Specificity	88.57%

Test Data Metrics

Accuracy	88.65%
Sensitivity	89.64%
Specificity	88.06%

6. Conclusion

    The Model seems to predict the Conversion Rate accurately and we would be able to reach the target of 80% conversion with the help of this Logistic regression model.

    The Top 3 Feature variables causing conversions are-

- Tags Closed by Horizzon
- Tags Will revert after reading the email
- Lead Source Welingak Website