Description Lead scoring – Case study

X Education, an education company, sells online courses to industry professionals. The company gets leads through different sources , Only few of them get converted. Current lead conversion rate is around 30% . The company want logistic regression model to be built and generate a score for the leads, so that the conversion rate increase to 80%.

OBJECTIVE:

1. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used to get target potential leads.
2. Current lead conversion rate is around 30% present target is 80%.

Solution Approach

The solution approach are as follows

1. Data loading and understanding
2. Data Pre Processing
3. Exploratory Data Analysis (EDA)
4. Data Preparation
5. Model Building and Evaluation

The dataset

The dataset have 37 feature and 9240 points..There are null value in the dataset those need to be taken care. The data type are in e proper format except 'total visit', 'Asymmetrique Activity Score', 'Asymmetrique Profile Score'. The data type can be change to int in place of float.

Data Pre Processing

The column having null value more than 40% were dropped , also The column having more than 40% value 'Select' including null value were dropped . Column having invariance data were dropped. The rows containing null in different column were dropped.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was carried out for univariate, bivariate and multivariate features. Plot and summary was given in presentation.

Data Preparation

- Data preparation was carried out for model building. Binary variables (Yes/No) were Converted to 0/1. Prospect ID and Lead Number no use in the analysis, so dropped. Category columns having more than 2 value was dummified. Data set was split in to train and test set at 70:30 ratio. The numerical column were scaled by min max scalar.

Model Building

- **First logistic regression model was build using stats model library and significance value was checked. Significance value more then 0.05 was dropped one by one . Then model was refitted. Variance_inflation_factor was checked and kept below 5.**

Model Evaluation

- **Train and test accuracy was calculated using sklearn matrix.**
- **Confusing matrix was calculated.**
- **Sensitivity and specificity were calculated**

Results

- **Logistic regression model was build wit 96% accuracy.**
- **1.Tags_Closed by Horizzon, 2. Total Time Spent on Website 3. Last Activity were the most important features recorded**
- **Sensitivity and specificity was Calculated to be 96 % and 97 % respectively**
- **Score was assigned to the leads.**
- **Leads with highest score should be contacted in priority basis for success conversion.**

```
# Predicted          Not_Converted    Converted

# Actual

# not_Converted       2059                  72

# Converted             81                   1754
```

Lead Score

	Converted	Converted_Prob	Prospect ID	predicted	Score
2495	1	99.991509	2495	1	99.991509
4123	1	99.969178	4123	1	99.969178
5293	1	99.967090	5293	1	99.967090
1991	1	99.962366	1991	1	99.962366
818	1	99.955040	818	1	99.955040
6944	1	99.954404	6944	1	99.954404
1803	1	99.936533	1803	1	99.936533
3739	1	99.933271	3739	1	99.933271
4613	1	99.912624	4613	1	99.912624
4252	1	99.911669	4252	1	99.911669
2984	1	99.906565	2984	1	99.906565
5372	1	99.903884	5372	1	99.903884
3246	1	99.902981	3246	1	99.902981
4892	1	99.902851	4892	1	99.902851
7098	1	99.899595	7098	1	99.899595
3290	1	99.899461	3290	1	99.899461
7202	1	99.897859	7202	1	99.897859