Lead Scoring Case Study DS C44

Summary

Submitted by:

1. K Chandrahas, chandrahaskuridi007@gmail.com , 7090858334
2. Thakur Yashraj Singh, yashrajsingh2301@gmail.com, 7207846570
3. Jay Prakash George, jgjaygeorge@gmail.com , 9713471908

Scope:

The case study is focused on developing a model for X Education to improve their lead conversion rate. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.

Objective:

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Methodologies:

• **Data Understanding, Preparation and EDA**

- **Data Dimensions**: (**9240 rows, 37 columns)**
- **Data Imputation** :
  - 'Select' values were replaced with 'Null'.
  - Values in some columns with 'Null' values were imputed with 'Others', 'Not Sure', etc. on a need-to-need basis.
  - Values in some columns which would result in a lot of unnecessary dummy variables were clubbed together and imputed with 'Others'.
- **Null Value Handling** :
  - Columns with more than 70% null values were dropped.
- **Data Imbalance** :
  - Columns with High Data Imbalance were dropped from the analysis.

- **Dummy variables** were created for the categorical features.
- Data was divided into two parts for **Dependent & Independent variables** (on the basis of column '**Converted**').
- Data was split into **Train**&**Test** sets (random 70:30 split).
- Data Normalisation/Scaling was performed using **MinMaxScaler.**
- **Feature Selection**:

  - **RFE (Recursive Feature Elimination)** : 15 variables were selected using RFE.

  - **Manual Feature Selection** : Logistic Regression was fitted on the Train data and p-values & VIF were calculated for each feature. Elimination happened recursively until we reached acceptable p-value & VIF for all the features.
- **Model Evaluation** (Cut off value of 80% was used)

Train Data:

- Accuracy : **91.1%**
- Sensitivity : **79.7%**
- Specificity : **98.1%**
- ROC Curve value **: .98**

Test Data:

- Accuracy : **90.9%**
- Sensitivity : **79.3%**
- Specificity : **98.04%**

- ROC Curve Value **: .97**

Conclusion & Recommendations

The Model seems to be predicting the **Lead Conversation** very well, with a good balance of **all evaluation metrics** & good **ROC Curve** value.

- The Sales team can leverage this model with **high confidence** to call the leads.
- If the need arises for a model with **higher Sensitivity** (to make calls to only most potential leads), we can **increase** the **cut-off** even more.
- If the Sales team decides to **expand their Leads coverage**, **cut-off** can be **decreased** to meet the demands.