Lead Scoring

X Education

Objective

- Building Logistic regression model & assigning Lead Scores to the prospective candidates of XEducation

- **Problem description**

- X Education is an online Education company which has Lead database, some of which got converted & some didn't
- The typical lead conversion rate is 30% which is expected to be maximized to atleast 80%
- Target is to identify the 'Hot Leads' which have a high conversion rate.
- The 'Hot Leads' to be identified by cutoff Lead Scores
- Lead scores to be assigned to each candidates based on probabilities calculated by Logistic regression model

Contents

- Data Inspection and Missing value treatment
- Dummy variable creation
- Logistic regression modelling
- Model Accuracy Check
- Model fit on test data
- Conclusion
- Recommendations

Data Preparation

- ❑ **Data Inspection and Missing value treatment**
  - ○ Columns containing >70% missing data were dropped.
  - ○ 'City' column had ~40% missing values & was dropped
  - ○ In absence of any visible correlation with Activity & Profile, these columns were dropped too
  - ○ *Asymmetric Index columns were checked for any possible relation to impute missing values
  - ○ Other columns with possible imputations were handled appropriately
- ❑ **Unique value columns**

  Columns with only one type of unique values were dropped in absence of variability
- ❑ **Imputation**

  High missing value containing columns were imputed with suitable values

Cleaned dataset

<table>
<tr><th>LeadOrigin</th><th>LeadSource</th></tr>
<tr><td>Do Not Email</td><td>Converted</td></tr>
<tr><td>TotalVisits</td><td>Total Time Spent on Website</td></tr>
<tr><td>Page Views Per Visit</td><td>Last Activity</td></tr>
<tr><td>Country</td><td>Specialization</td></tr>
<tr><td>Tags</td><td>Lead Quality</td></tr>
<tr><td colspan="2">What is your current occupation</td></tr>
<tr><td colspan="2">A free copy of Mastering The Interview</td></tr>
<tr><td colspan="2">Last Notable Activity</td></tr>
</table>

- After data cleaning, 15 columns were left.

Outlier treatment

- Numeric columns were treated for Outliers
- Data within +/- 3*Standard deviation was retained

Dummy variables & Numerical encoding
Train-Test split

- To start with Logistic regression, Dummy variables are created with Original Categorical variables being dropped after dummy creation.
- Yes & No values in columns are converted to 1 &0 respectively. The final dataset contains-

1. Rows:9112
2. Columns:150

- Final dataset is split into train and test dataset in 70%-30%proportion.
- Train & Test data are split into X &Y.
- Y is taken as 'Converted', remaining variables as X

Model Building

- 15 Features were selected using RFE.
- Six Logistic regression models were built iteratively
- Final model was selected based on:

1. p-values < 5, indicating absence of multicollinearity

- Model performance measures

1. High values of Accuracy, Sensitivity & Specificity indicate good predictive powers of model.
2. Low False positive rate indicates model's ability to predict positive values accurately.

Measures

Accuracy	89.21%
Sensitivity	83.35%
Specificity	92.92%
False Positive Rate	7.07%
Positive Predictive Value	88.20%
Negative Predictive Value	89.79%

Model Accuracy Check

- Accuracy, Sensitivity& Specificity plot to find optimum cutoff for probability
- The three curves intersect at ~0.32.
- Model accuracy at this point is 89.21%, which is very close to earlier calculated value

Model fit on test data

- Final model was fit on the test data.
- Predictions of Converted values were made.
- The accuracy achieved on test dataset is also same at 88.84%.
- Sensitivity of 83.13% and Specificity of 92.30% was achieved.
- These measures indicate a good fit of model on the test data as well

Conversion

- To calculate Conversion on the entire dataset, a master data frame was created with final y(s) from train and test sets.
- From train, 'y_train_pred_final' and from test, 'y_pred_final' are concatenated
- Cutoff Lead Score was applied on this dataset to select only Hot leads
- At Lead Score of 38, Conversion of 88% was achieved, which is more than target of80%

Recommendations

- To get more customers, X Education must keep the lead score lower, starting at '0'.
- But to achieve target conversion of greater than 80%, it should keep the cut off at 30.Thus, in the model, data frame changed for cut off
- Lead Score to gauge the Conversion percentages w.r.t. actual converted.
- Lowering the lead score cut off reduces conversion %, but it increases number of actual converted.
- Based on the man power availability with X Education, it may decide to give weightage to conversion % or actual numbers.