Lead Scoring case study

SUMMARY REPORT

X-Education, an online education company is facing a poor lead conversion rate. Out of 100 leads a day,30 is getting converted to payments. To make this process more efficient, the company wishes to identify the most potential leads.

The case study aims to identify promising leads and thereby, increase the lead conversion rate.

The following steps were involved in the analysis.

Data reading and understanding:

- ✓ *Importing the required libraries like numpy, pandas, matplotlib, seaborn etc.*
- ✓ *Reading the dataset to 'leads' dataframe. Basic inspection is done.*
- ✓ *Dataset consists of 9240 rows with 37 columns.*
- ✓ *Few aspects like null values, data types of all variables; and statistical information of continuous variables were checked.*

Data Cleaning and Preparation:

- ✓ *Few unwanted columns like ID columns, system and sales team generated columns were removed.*
- ✓ *Some Categorical variables had 'Select' option which were imputed as missing values.*
- ✓ *Columns with more than 40 % missing values were removed/dropped.*
- ✓ *Null values were imputed/dropped appropriately so that not much of data was lost.*
- ✓ *Categorical variables which were highly skewed towards one category was dropped.*
- ✓ *Categories will very less count (0.0001%) were clubbed together in 'Lead_source' variable.*
- ✓ *Incorrect values in the variable were corrected.*
- ✓ *Outliers were dropped.*

EDA:

- ✓ *Univariate analysis for continuous and categorical variables were performed using distribution plot and count plot, respectively. A function was defined to perform EDA on all columns*
- ✓ *Bivariate analysis for the categorical variables w.r.t target variable 'Converted' was performed using count plot.*
- ✓ *Bivariate analysis for continuous variables was done using scatter plots.*
- ✓ *Heatmap was plotted to identify the correlation between target variable and other numerical variables.*

Creating Dummy variables:

- ✓ *The categorial variables with Yes/No options were mapped and converted into binary (0/1).*
- ✓ *Dummy variables were created for the categorical variables with multiple categories using one hot encoding. The original variables were dropped afterwards.*

Test-Train Split

- ✓ *The dataset was divided into train data and test data on 70:30 ratio. Each dataset was further divided into target variable and feature variables.*

Scaling

- ✓ *The numerical variables were scaled using MinMaxScaler, so that the variables are within the range zero and one.*

Dummy variables correlation:

- ✓ *The correlation between dummy variables were identified using a heatmap and the highly correlated variables were dropped.*

Model Building:

- ✓ *Out of 42 variables ,15 variables were selected using RFE for model building.*
- ✓ *By checking the p value and VIF values, insignificant variables were dropped, and model was rebuilt.*
- ✓ *Rebuilt the model until all the variables in the model were significant and had no multicollinearity.*
- ✓ *Using the final model, the target variable values were predicted.*
- ✓ *Compared the actual values and predicted values*

Accuracy, Specificity and Sensitivity

- ✓ *Accuracy of 79.2% was obtained from the confusion matrix with the predicted values.*
- ✓ *Specificity and sensitivity were calculated to verify the correctness of final model.*

ROC Curve

- ✓ *The roc curve was plotted. True Positive Rate and False Positive Rate to verify the accuracy of the model.*
- ✓ *The curve follows the left-hand border and then the top border of the ROC space, hence the test is accurate.*

Optimal Cutoff Point

- ✓ *The Accuracy, specificity and sensitivity were calculated for different probability cutoffs.*
- ✓ *The optimal cut off was found out to be 0.29 from the line plot between accuracy, specificity, and sensitivity.*

Precision-Recall Curve

- ✓ *A precision-recall curve was plotted to find the relationship between precision (= positive predictive value) and recall (= sensitivity) for every possible cut-off.*
- ✓ *Precision score on Train Data = 0.7034*
- ✓ *Recall Score on Train data = 0.7874*

Prediction on Test data

- ✓ *Predictions were made on the test data using the model built.*
- ✓ *Accuracy, Specificity, Sensitivity, Precision and Recall were calculated for the test data and the values were compared with the train data.*
- ✓ *Precision score on Test data = 0.7110*
- ✓ *Recall score on Test data = 0.7750*

Lead Score Calculation

- ✓ *Lead Score were assigned to customers to identify hot leads (leads that are most likely to be converted).*
- ✓ *This was calculated by multiplying conversion probability with 100*
- ✓ *From lead score, it was inferred that higher the score, higher are the chances for a lead to convert*
- ✓ *A score of 60% and above is desired.*

Conclusion

- ✓ *The top variables which contribute the most towards the target variable were identified.*
- ✓ *The variables which were inversely proportional to the target variable were also identified.*

The following variables are directly proportional to the 'Converted' variable

1. *Total Time Spent on Website*
2. *Lead Origin_Lead Add Form*
3. *Lead Source_Welingak Website*
4. *Total Visits*
5. *Lead Source_Olark Chat*

The following variables are inversely proportional to the 'converted' variable

1. *Lead Origin_Landing Page Submission*
2. *Specialization_Others*
3. *What is your current occupation_Other*
4. *what is your current occupation_Student*

5. *What is your current occupation_Unemployed*
6. *Specialization_hospitality Management*

Challenges Faced during the analysis

- *Some of the categorical variables had the default option 'select' as a value. Those values had to be replaced with null values.*
- *Many categorical variables were skewed, i.e., majority of the data belonged to one category. Had to drop the variables to reduce the complexity of the model.*
- *Some of the categorical variables had a lot of sub-groups/categories with very less percentage of count. Those categories had to be clubbed together to reduce the complexity of the model.*

Information learned:

- *Customers who spent more time on website are more likely to convert. (Time Spent on Website)*
- *Customers whose lead origin is lead add form are more likely to convert than other categories.*
- *Customers whose lead origin is Landing page Submission are less likely to convert.*
- *Student and unemployed individuals are less likely to convert.*
- *The more times the customer visit the website, the more likely he is to enrol. (TotalVisits)*
- *Customers whose lead source is Olark chat seems to have converted a lot.*
- *Customers with specialization as 'Others' hardly enrols.*
- *Customers whose source is Welingak Website are likely to converts more.*
- *A LEAD SCORE OF 60% or MORE IS DESIRED.*

DONE BY
ANUPAMA RAJEEV
SUMITHA T

----------*** END *****----------**