SUMMARY

LEAD SCORING CASE STUDY

What is the problem?

➢An education company – X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. People visits the website, fill up the required details, search the courses, such people are classified to be a lead. Once these leads are acquired, employees from the sales team start contacting them via different mediums. Through these process, some of the leads get converted while most do not. The typical lead conversion rate at X Education is around 30%.

What the company wants?

➢The company wants, the leads that are most likely to convert into paying customers. The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO has given the target lead conversion rate to be around 80%.

Data provided:

➢We have been provided with a leads dataset (Leads.csv) from the past with around 9000 data points. This dataset consist of various attributes such as Lead Source, Total time spent on website, Total visits, Last activity etc.

Goals of the case study:

➢Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads in the following steps:

Step 1: Reading and Understanding the data-

- After importing the required libraries, read the given csv file ‘Leads.csv’.
- Checked the number of rows and columns (9240, 37) and datatypes of each columns.

Step 2: Data Cleaning-

- Checked the data to be balanced or not in each columns.
- Checked for the null values and dropped the columns with more than 45% of missing values such as, ’How did you hear about X Education’, ‘Lead Quality’, ‘Lead Profile’, ‘Asymmetric activity index’, ‘Asymmetric profile index’, ‘Asymmetric profile score’.
- After dropped these columns, shape of the data (9240, 30).

Step 3: EDA-

- CATEGORICAL ANALYSIS OF DATA- Columns such as lead origin, lead source, last activity, country, specialization, what is your current occupation, tags, city, a free copy of mastering the interview, last notable activity are firstly analyzed univariate analysis then analyzed by bivariate analysis with converted.
- NUMERICAL VARS ANALYSIS OF DATA- Columns such as lead number, converted, total visits, total time spent on website, and page views per visit are analyzed with correlation.
- Handled outliers by capping.

Step 4: Created Dummy Variables-

- We went on with creating dummy data for the categorical variables.

Step 5: Test Train Split-

- The next step was to divide the dataset into train and test data with the proportion of 80%-20% respectively.

Step 6: Scaling-

- We used standard scaler to scale the original numerical variables.

Step 7: Feature Selection Using RFE-

- Using the Recursive Feature Elimination , we selected the top 15 important features to create a model.
- Later the rest of the variables were removed manually depending on the VIF values and P-values(The variables with VIF<2 and P-value<0.05 were kept).

Step 8: Model Evaluation-

- A confusion metrics was made and calculated the overall accuracy.
- We also calculated the Accuracy,Sensitivity ,Specificity and also plotted the ROC curve with 0.5 and above predicted values to be considered as 1 to understand how reliable the model is. We now tried to make it more optimized. We went ahead with the following steps

Step 9: Plotting The ROC Curve-

- We then plotted the ROC curve for the features and the curve came out be pretty decent with an area coverage of 93%.

Step 10: Finding The Optimal Cut Off Point-

- We plotted the probability graphs for the Accuracy, sensitivity and Specificity for different probability values. The intersecting point of the graph was considered as the optimal cut off point. The cut off point was found out to be 0.28.

Step 11: Precision And Recall Metrics-

- Based on the Precision and Recall tradeoff , got a cut off value of 0.34.

- The confusion matrix build with predicted values over 0.34 to be considered as 1 was further adopted.

- The obtained accuracy :87.32% , sensitivity:82.95 ,specificity:89.98% was noted and found that 0.34 was a better threshold value

Step 12: Predictions on Test Dataset-

- We implemented the learnings to the test model and calculated the conversion probability based on the sensitivity and specificity and found out the accuracy=87.80%, sensitivity =83.64%, specificity=90.42%.
- Also checked the precision and recall on the test dataset for the 0.34 threshold be 84.6% and 83.6% respectively and the ROC area under the curve be 94%.

RESULT-

- It seems to be a good model and ready to be presented in front of the CEO of X Education.
- X Education can be flourished as they focus on the variables that mattered the most in the potential customers such as spending time on the website, last activity, tags and origin of higher conversion lead add form
- Certain business goals were provided like automated email and SMS to be sent , contacting previous learners and references , calling only when time spent on website or visiting the site was more.