Lead Scoring Case Study

Problem Statement

- An Education Company named ‘X Education Company’ sells online courses to professionals.
- The Company markets its courses on several websites and search engines.
- The professionals who are interested in online courses visit their website and browse for the courses.
- Once they fill up a form in their website providing their contact details , then those members will be termed as “Leads”
- After acquiring these “Leads” company’s marketing team will focus on reaching these people to get them Converted.
- The Company’s current conversion rate is only 30% inspite of getting more leads.
- In order make this process efficient , the company wishes to identify “Hot Leads” who are the potential leads of X Education company.
- If this is successful, then Sales team can identify the potential leads and make calls only to them rather than reaching out to the people who are less likely to convert or take up a course.
- By this method they can surely improve the Conversion rate of the company and also reduce the effort taken by the Sales team.

Case Study Approach

- Lead Score case study involves the prediction of potential leads or “Hot Leads”
- The data set contains 37 predictor variables
- To begin with, the python libraries were imported and the data frame was created by uploading the csv file
- The insignificant variables were dropped as a first step followed by missing value treatment
- The columns with missing values greater than or equal to 40% were dropped
- For few of the categorical variables , missing values were imputed with ‘Mode’
- Some of the Categorical variables with Yes and No as values were mapped to 1s and 0s
- For the categorical variables with more than 2 levels , dummies were created
- Once the dummies were created the repeated variables were then dropped from the data frame
- The data frame was then inspected for any Outliers
- After all the data processing steps, the data frame was split into Train and Test sets using Scikit-Learn library
- Then the variables were subjected to ‘Feature Scaling’ to ensure they are in same range or scale, followed by checking the imbalance rate of the data set

Case Study Approach Contn’d

- The imbalance rate was pretty decent around 38% and hence there were no techniques employed to handle it
- A heat map was produced using Seaborn library and the most correlated variables were listed
- Th highly correlated variables were dropped for both train and test sets
- The first Logistic Regression model was built using Statsmodel library using 81 predictor variables
- Post building our first model, automated feature elimination technique called “RFE” was employed, and 25 variables were selected
- The Second Model was built using the features selected by RFE, whose Accuracy came out to be 90%
- Based on the p-value and VIF value some of the insignificant variables were dropped
- After dropping the variables with VIF value greater than 5 , the third model was built and the Accuracy of the model was 88% and did not see a drastic dip
- Other parameters such as Sensitivity and Specificity were calculated , which came out to be 81% and 93% respectively
- The ROC curve was plotted to understand the trade off between Sensitivity and Specificity , and optimal cut-off was also measured
- Finally, a Model was built and predictions were made on the test set

Data Processing

- As a first step of Data Processing variables like ‘Prospect Id’, ’Lead Number’, ‘Country’ and ‘City’ were dropped as they do not contribute much to the prediction
- Few Categorical variables which had ‘No’ as response for all the records were dropped , as they wouldn’t make a significant contribution in the prediction
- Missing values were then handled by dropping some columns and imputing some values
- The Categorical variables like 'Do Not Email’, 'Do Not Call’, 'Search’, 'Newspaper Article’, 'X Education Forums’, 'Newspaper’, 'Digital Advertisement’, 'Through Recommendations’, 'A free copy of Mastering The Interview’ were mapped to Binary values 1s and 0s
- For the other categorical variables with multiple level , dummies were created and merged with the Original Data frame
- The Data frame was split into Train and Test set on 70:30 ratio
- After the split, Feature Scaling was performed on the variables, ‘TotalVisits’ , ‘Page Views Per Visit’ and ‘Total Time Spent on Website’ to make their scale aligned with other variables
- The Data Frame was examined for the Imbalance ratio , which came out to be 38% , hence the Converted Rate is 38% for the data set

Finding the Correlations

- A Heat map was created to find the highly correlated variables
- The following highly correlated variables like 'Newspaper Article', 'X Education Forums','Direct Traffic', 'Facebook', 'Act_Email Bounced', 'Act_Email Link Clicked', 'Act_Email Opened', 'Act_Email Received','Act_Had a Phone Conversation', 'Origin_Lead Import','Origin_API', 'Act_SMS Sent','Act_Unsubscribed', 'LastAct_Email Received', 'LastAct_Modified', 'WeLearn','Flexibility & Convenience' were dropped from both Train and Test sets.
- After dropping the highly correlated variables, a heat map was produced again to ensure none of the variables have very high correlation coefficient
- Removing the highly correlated coefficients in the data set , is advisable to improve the efficiency and accuracy of the Model built
- Having these highly correlated independent variables will hamper the model prediction as, a change in one variable would cause significant change in the other related variable

Model Building

- After all the data processing steps , Splitting the data set and finding out the correlations , the first Logistic Regression model was built using “Statsmodel” library
- The p-value of the variables was 0 for almost all the variables, hence no variables were dropped based on the p-value
- The Coefficients for the variables ‘LastAct_SMS Sent’, ‘Reference’, ‘Total Visits’, ‘Total Time Spent on Website’ came out to be high showing that these variables may play a significant role in prediction
- For certain variables like ‘LastAct_Email Bounced’, ‘Tag_Closed by Horizzon’ and ‘Do Not Call’, although the coefficients are high, we cannot count on them for prediction as they may not be a good predictor variables. For instance, the members who have opted for ‘Do not call’ option may not be really interested in the course so counting on them may not fetch us an accurate result

Feature Elimination Technique

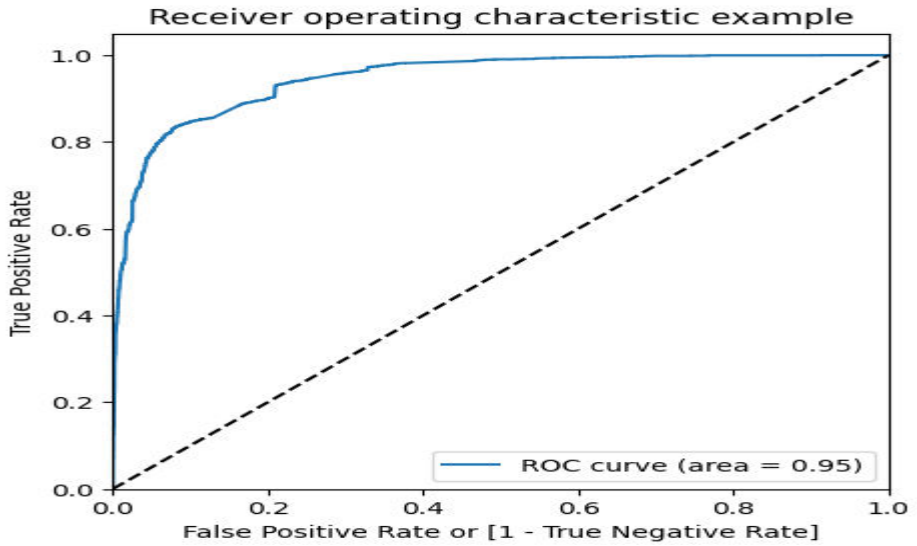
- To select the significant variables for Model prediction an automatic feature elimination technique called “RFE” was implemented
- RFE will rank the variables based on their significance
- Using RFE, 25 most significant variables were selected, and our second Model was built using the selected features
- Many variables had high p-values in the second model, which means the variables with high p-value are insignificant
- The VIF- Variance Inflation Factor was measured for the variables to check if there is any collinearity among the variables
- For the variables ‘Act_Email Marked Spam’ and ‘LastAct_Email Marked Spam’ , the VIF value was calculated as ‘Infinity’ which shows those are highly correlated and those variables were dropped
- The variables with VIF values greater than 5 were dropped and the final model was built

Building a Confusion Matrix

- A Confusion Matrix is built to gauge the Model performance
- There could be possibility that our Model misclassifies people as “Hot Lead” when they are actually not and vice-versa
- To avoid this and to measure the Model’s performance using various factors , confusion matrix is required
- Using the confusion matrix, we calculated the Final Model’s Accuracy as 88%
- The Sensitivity and Specificity which measures the True Positives and True Negatives came out to be 81% and 93% respectively
- Since the metrics came out to be good and VIF values were very low, we did not drop any more variables
- The third model is considered as the Final model, and we started with the Prediction

ROC Curve

- A ROC Curve was plotted to understand the tradeoff between the Specificity and Sensitivity
- The AUC(Area Under the Curve) value is measured for the model, the AUC came out to be 0.95 for the Model
- Higher the AUC value, higher will be the Model performance or Prediction power
- Using ROC, we can find out the optimal cut-off and the model's performance can be calculated for the optimal cut-off obtained.
- We got optimal cut-off value as 0.4 and the metrics like Accuracy, Sensitivity and Specificity were measured which did not have major difference



Making Predictions

- Feature Scaling was done on the Test set and same number of features were fed into Test set
- Using ‘Predict’ framework, predictions were made on the test set
- A confusion matrix was then built, and the Accuracy, Sensitivity and Specificity was calculated for the test set
- Metrics of Test Set are as follows

  Accuracy : 88%

  Sensitivity: 80%

  Specificity: 93%

- We did not see a major difference in the performance of the Model on train and test sets , hence we can conclude this is a good model
- Finally Lead Scores were assigned to the members , by multiplying the Conversion Probability into 100

Recommendations

- RFE method can be used iteratively to select/eliminate features and Models can be built for each set when there are greater number of independent variables
- VIF and p-values can be combinedly used to eliminate insignificant variables
- The variable ‘LastAct’ seem to have a greater coefficient and can be considered as a significant variable for prediction
- The variable ‘Tag’ has mostly negative coefficients , hence ‘Tag_Busy’, ‘Tag_Alread a Student’ can be classified as insignificant ones and the Sales team need not focus on those members
- On the other Sales team should reach out to people whose value is yes for variables ‘LastAct_Had a phone conversation’, ‘LastAct_Sms Sent’, ‘LastAct_Email Opened’ as these people are more likely to take up the course
- The Sales team should focus on reaching the people whose “Time Spent on website” is high as these people must be very much keen to take up the course.
- People who often visit the website and spend time might be investigating on the courses so team should reach them frequently to clear their doubts and make their way of choosing the courses easier.
- Also, people visiting the website often should be promised with offers once they take up the course within a particular date , this will enable them to take faster decisions without oscillating or looking for other platforms

Recommendations

- The team can refrain from making phone calls if the value is ‘1’ for the dummy variables LastAct_Unsubscribed,LastAct_Email Bounced,Tag_opp hangup,Tag_Not doing further education,Tag_Graduation in progress etc., as they clearly indicate that probability of conversion would be very less , instead they can concentrate on variables like Tag_Interested in other courses, Tag_Interested in full time MBA, Tag_Shall take in the next coming month as there is high probability of lead conversion if these variables have value as ‘1’ or ‘Yes”
- The Lead Score can be used as a important criteria to identify the potential members and only those Leads can be contacted to minimize the efforts of the Sales team and also getting a good conversion rate
- The leads whose Lead Score is greater than 90 could be classified as Hot leads , thus the high conversion rate can be obtained

Conclusion

- The Logistic Regression model was thus built by analyzing and processing the data set
- Hybrid method like RFE,VIF and p-value were employed to build a Model with high Accuracy and other parameters.
- Finally, the predictions were conducted on the test set and hot leads were identified by the Model