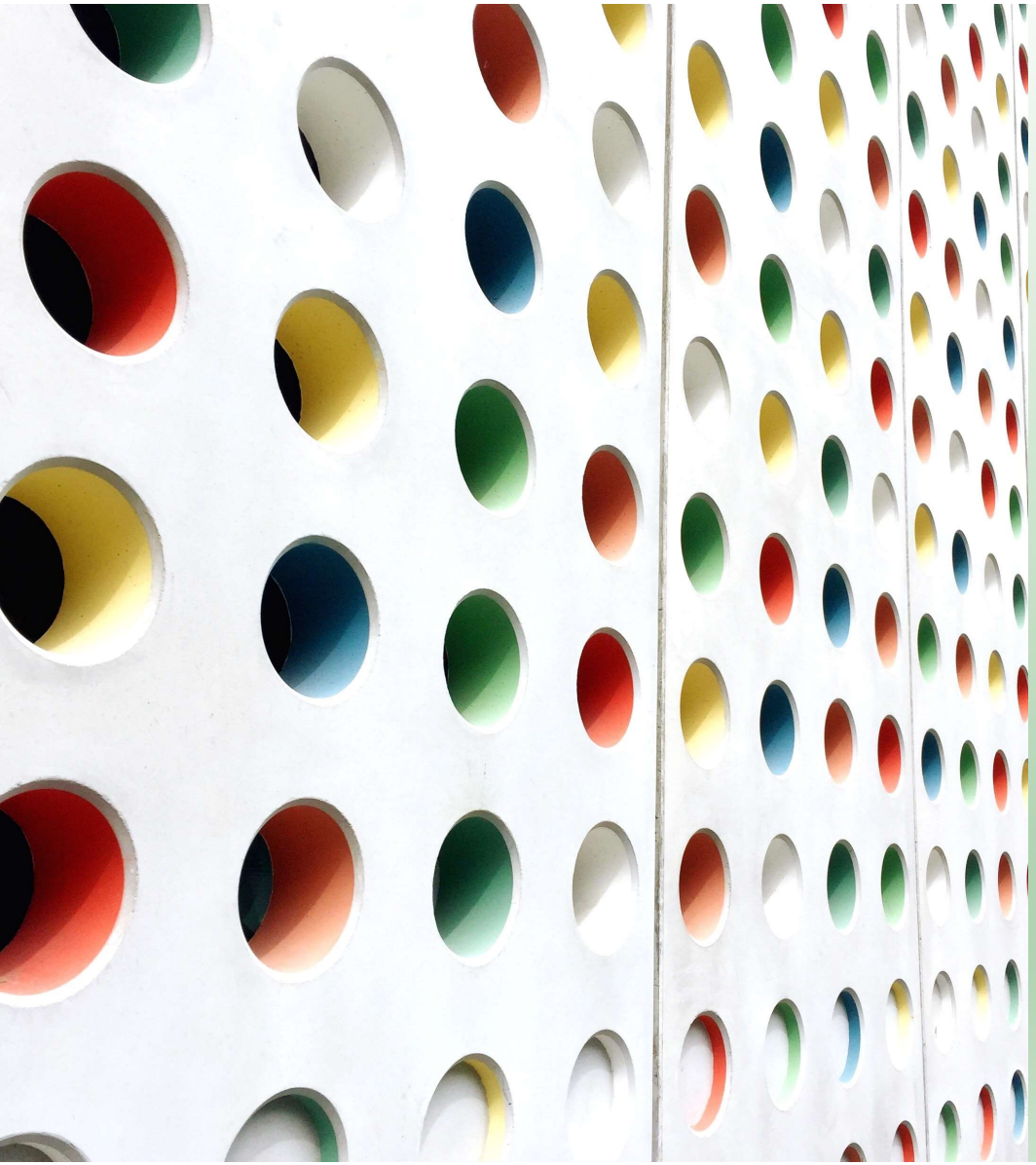

Lead Scoring Case Study

SUMMARY BY

BIJI KRISHNA , BAPPI BANIK, SANTANU BISWAS

Problem Statement

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Business Goal

Although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Solution Summary

Step 1 : Import Libraries and Data : Imported all required libraries and data steps
Step 2: Inspect Data : Check head , missing values and duplicates
Step 3: Clean Data : Replaced Select with Nan values, Encoded the variables with Yes/No labels with 1/0
Step 4: Handling Missing values.Dropped columns having more than 30% missing values. Dropped the rows for the rest of the missing values which is less 1.5%
Step 5: Exploratory Data Analysis
Plotted different features Vs Converted column to understand the Conversion rate and impact. Used Boxplot or count plot or subplot to analyse different parameters. Renamed a few columns where there are not enough data available
Step 6: Created Dummy Variables and Dropped original columns for which dummy is created
Step 7: Test_Train Split - Assign Feature Variables to X and Target Variable to y
Step 8: Feature Scaling - Scaling features which are not 1 and 0

Step 6: Created Dummy Variables and Dropped the original columns for which the dummy is created
Step 7: Test_Train Split – Assign Feature Variables to X and Target Variable to y
Step 8: Feature Scaling – Scaling features which are not 1 and 0
Step 9: Analyzing Correlation – Check Conversion Rate 44.22335025380711
Step 10: Building Model – with all feature
Step 11: Feature Selection done Using RFE – Selecting 15 Features, then made a model with RFE selected top features. As per the VIF and P-Values dropped a few features and also removed highly skewed features. Got the predicted values on the train set.
Step 12: Assessed the model with StatsModels
Step 13: Checked VIF
Step 14: Created Dataframe with True Conversion and predicted probabilities
Step 15: Metrics like Accuracy, Sensitivity, Specificity, Precision Score, and Recall were checked
Step 16: Plotted the ROC Curve
A ROC curve demonstrates several things:
It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity). The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test. We got an approx. area under the curve 86%
Step 17: Finding the Optimal Cutoff Point
Step 18: Plotting Accuracy, Sensitivity and Specificity for various probabilities
From the Accuracy Sensitivity and specificity above, 0.38 is the optimum point to take it as a cutoff probability.
Step 19: Precision and recall tradeoff
Step 20: Making predictions on the test set- checked if 80% of cases are correctly predicted based on the converted column.
Step 21: Accuracy, Sensitivity and Specificity values of the test set are around 78%, 79% and 77% which are approximately closer to the respective values calculated using the trained set.
Step 22: Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 0.799074686054197 (train) / 0.7954887218045112 (test}. Hence overall this model seems to be good.
Step 23:The final model has a Precision of 0.74, this means 74% of predicted hot leads are True Hot Leads
Final Prediction conversion on both train and test set is around 80%+ which is in line with the target.
Overall this case study helped to understand the details about an education company's lead generation process and how to improve the conversion rate.


Thank You