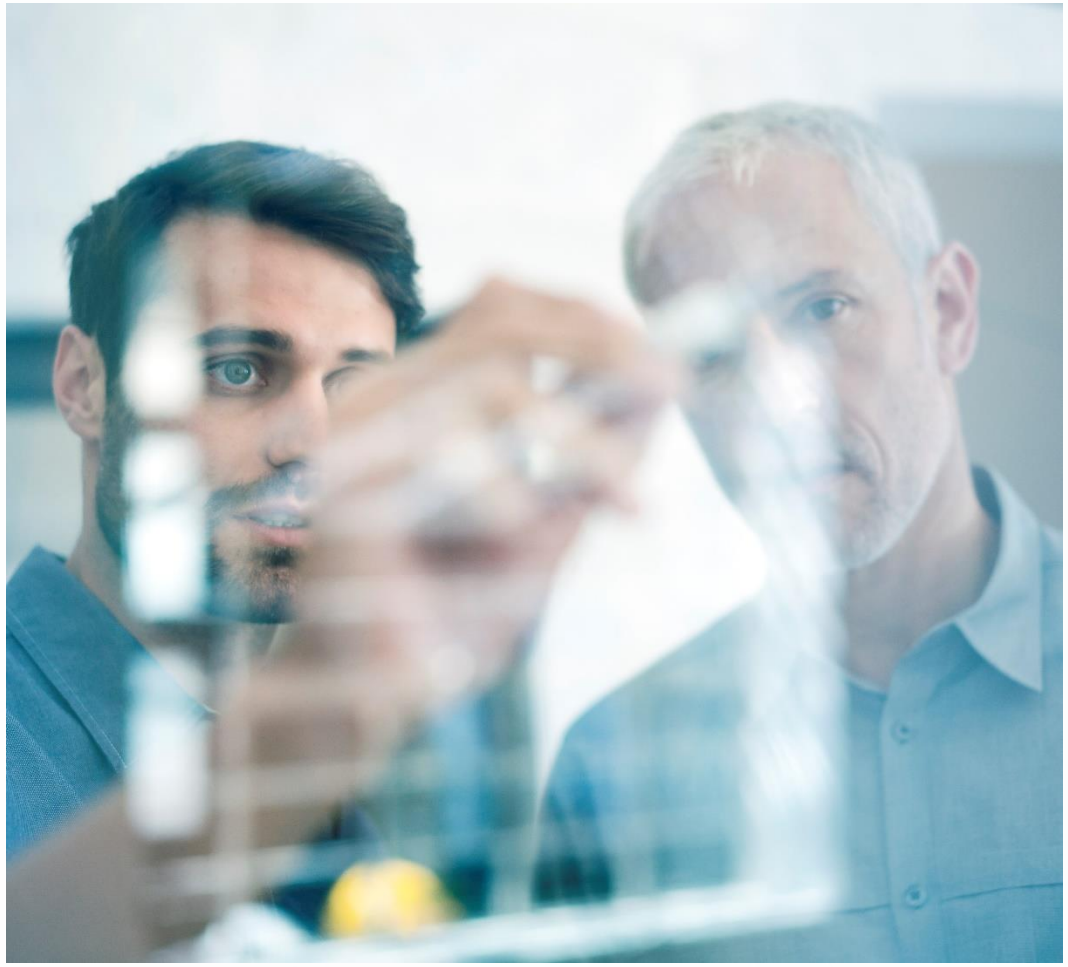


X Education business plan

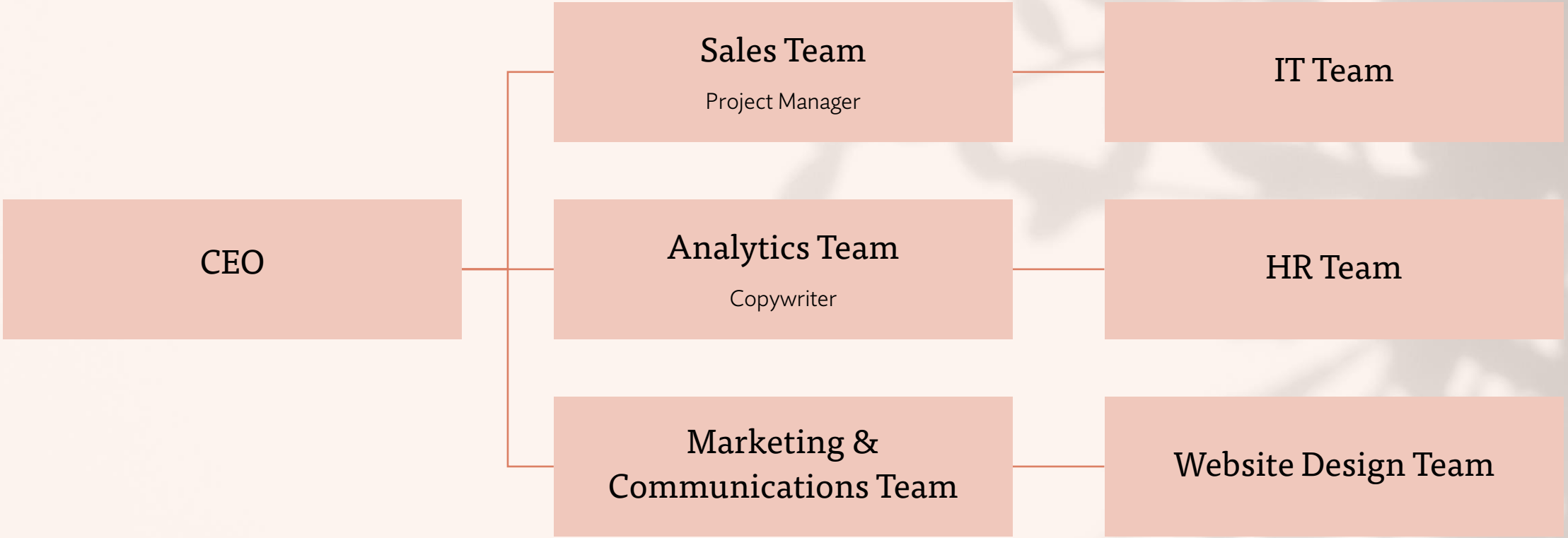
Divya Sonali Minz

Problem statement

X Education sells online courses to industry professionals. The company gets lots of leads, but the conversion rate is very poor. the company wishes to identify the most potential leads, also known as ‘Hot Leads’. The sales team would then focus only on the hot Leads to increase the conversion rate from the current 30%.

The company requires to build a model which will help the conversion rate to be around 80%.

The Team Assumption



Objective

- The objective is to optimize lead scoring method based on characteristics like buying behaviour, buying tendency, fit and demographics. This can be done by implementing the lead scoring modelling which can be either explicit or implicit with the lead point system.

Goals of the Case Study

1. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
2. There are some more problems presented by the company which the model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well.

Analysis Approach Adopted

1. Sourcing of data
2. Analysing data
3. Cleaning data
4. EDA
5. Feature Scaling
6. Split the data into 2 datasets – test and train
7. Data preparation for modelling
8. Building of model
9. Evaluation of model – specificity and sensitivity i.e. precision recall
10. Run predictions on the test set

Sourcing, Cleaning and Data Preparation

1. Read the data from CSV file
2. Fix the outliers
3. Clean the data by managing null values and removing higher null values data
4. Identifying the redundant columns and removing those
5. Impute the null values
6. Data analysis exploration with conversion rate of approx. 38%
7. Standardization of features

Outliers

1. Total visit on the website
2. Total time spent on the website
3. Total number of views in each page per visit

Outliers

1. Total visit on the website
2. Total time spent on the website
3. Total number of views in each page per visit

Data Analysis and Data Conversion

1. Binary variables into 0 and 1
2. Created dummy variables for categorical variables

Feature Scaling

1. Split train and test sets
2.Feature scaling of numeric data
3. Split data into train and test sets

Model Building

1. Feature selection using RFE
2. Determined optimal using logical regression
3. Calculated accuracy, sensitivity, specificity, precision, Recall & evaluate model

Impact of Variables

1. Conversion rate
2. Total visits
3. Total time spent on website
4. Lead Score_Olark Chart
5. Lead Origin_Lead Add Form
6. Lead Score_Wellingak Website
7. Do not send email
8. Lead Score_Referral Site

Model Evaluation – Sensitivity and Specificity on Train Data Set

1. Graph represents optimal cutoff of 0.37 based on Accuracy, Sensitivity, Specificity
2. Accuracy = 78%
3. Sensitivity = 82%
4. Specificity = 76%

Model Evaluation – Precision and Recall on Train Data Set

1.Graph represents optimal cutoff of 0.42 based on Precision and Recall

2. Precision = 79%

3. Recall = 65%

Model Evaluation – Sensitivity and Specificity on Test Data Set

1. Accuracy = 78%
2. Sensitivity = 80.8%
3. Specificity = 76.5%

Result

1. Accuracy, Sensitivity, Specificity values of training and tests set are close to training set
2. Accuracy, Sensitivity, Specificity values of training set are 79%, 82% & 76% respectively
3. Accuracy, Sensitivity, Specificity values of test are 78%, 81% & 76% respectively
4. Conversion rate for train and test data set is 82.7% & 80.8% respectively
5. Prediction on the tests set is done using cutoff threshold from sensitivity and specificity metrics

Conclusion

1. While sensitivity-specificity and precision-recall metrics are checked, consideration has been made for optimal cutoff based on sensitivity and specificity for calculation of final prediction
2. Accuracy, sensitivity, specificity are approx. 78%, 81% & 76% respectively which are almost closer to Values calculated using Trained Data Set
3. Lead Score Calculated for the conversion rate final model on Train & Test dataset is 82.7% &80.8% respectively

Summary

1. Good number of leads have been generated at the initial stage i.e. top but only a few of them come out as paying customers from the bottom stage.
2. At the middle stage it is important to nurture the potential leads by communicating and timely information sharing on the benefits to the leads for better conversion rate.
3. The most important is to find out and segregate the best prospects out of the total leads generated.
4. Factors like 'Total Visits' , 'Total Time Spent on Website' , 'Page Views Per Visit' contribute towards higher conversion.
5. Maintain record of important leads so that they can be timely informed on the benefits like new courses, services, job offers and future higher studies.
6. It is key aspect to monitor each lead specifically to send tailored information based on their interest on the courses that suits to them.
7. Planning to capture the leads and convert them is vital.
8. Once the leads are converted, focus must be on them, their needs and interests.
9. Communication sessions can be organized to elicit and understand the needs of important leads and also to answer their queries on timely manner.
10. Follow up plan must be prepared and action must be taken to understand the behaviour and intensions of the leads with objective to convert as customers to join the online courses.

Thank you

Divya Sonali Minz

divyaminz@gmail.com