Introduction

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. 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%.

Goals

- 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.

- X Education has a period of 2 months every year during which they hire some interns. The sales team, in particular, has around 10 interns allotted to them. So during this phase, they wish to make the lead conversion more aggressive. So they want almost all of the potential leads (i.e. the customers who have been predicted as 1 by the model) to be converted and hence, want to make phone calls to as much of such people as possible. We have to suggest a good strategy they should employ at this stage.

- Similarly, at times, the company reaches its target for a quarter before the deadline. During this time, the company wants the sales team to focus on some new work as well. So, during this time, the company's aim is to not make phone calls unless it's extremely necessary, i.e., they want to minimize the rate of useless phone calls. Suggest a strategy they should employ at this stage.

Data preprocessing

Data cleaning

- In data the 'Select' string were present and made these as NULL values for analysis.
- Columns with > 52% of missing values were dropped.
- For some of categorical variables missing values are imputed using the string 'Not sure' or using the mode value.
- For numerical variables, if the percentage of data missing is very less dropped those records.
- Handles the outliers in numerical variables and capped and floored the data using 95% and 5% quantiles.
- Standardized the data

EDA

- Checked the data imbalance in the target variable.
- Performed univariate analysis and dropped the insignificant variables.
- Performed bivariate analysis and selected only the significant variables which are affecting the target variable.

Data preparation

- Converted binary categorical variables to numerical.

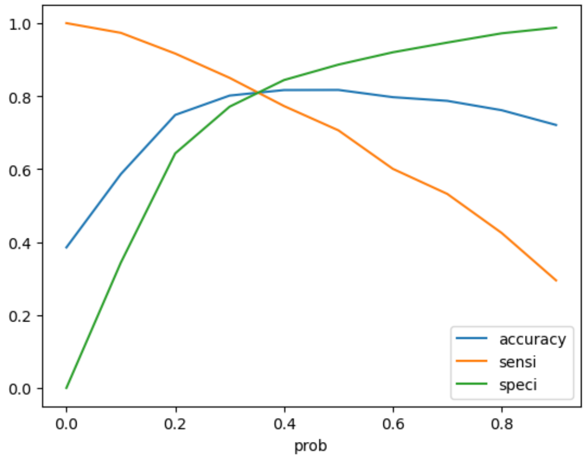
- Created dummy variables for few categorical variables.
- Split the data into train and test in the ratio 70:30.
- Scale the features using standardization

Model building

- Imported the model logistic regression
- Used RFE for feature selection – to reduce the feature to a smaller number from 47 to 20.
- Total 9 models were built to get a stable model with less p Value and VIF < 5.

Model Evaluation

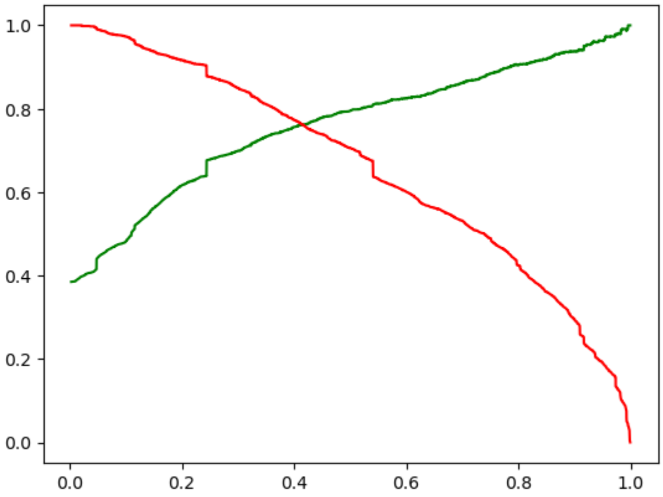
- Using ROC curve found out the optimal threshold as .35.



- For training set the KPI s were as below

  Accuracy 81.08%
  Sensitivity 81.72%
  Specificity 80.69%
  Precision 79.54%
  Recall 70.6%

- There is a trade-off between precision and recall as below. Hence we can chose sensitivity-specificity for test set.



- **For test data the KPI s are as below,**

  Accuracy 80.49%

  Sensitivity 80.48%

  Specificity 80.50%

  We can see that both training and testing data set have similar values for Accuracy, sensitivity and Specificity.

Feature Importance

- Key features are given below, with their parameter values

Lead Source_Welingak Website	5.811465
Lead Source_Reference	3.316598
What is your current occupation_Working Professional	2.608292
Last Activity_Other_Activity	2.175096
Last Activity_SMS Sent	1.294180
Total Time Spent on Website	1.095412
Lead Source_Olark Chat	1.081908

The top three variables which contribute most towards the probability of getting converted the leads are

- Lead Source_Welingak Website
- Lead Source_Reference
- What is your current occupation_Working Professional

Recommendations

The X education can follow the below strategies during the intern hire periods.

- Based on the model analysis, the company should contact leads who are spending more time on 'Welingak website'. Also, they can spend highly on advertising in 'Welingak Website'.
- The company can contact the reference provided by the learners to increase the lead conversion rate. The company can provide discounts for learners who are providing reference that converts to lead.
- Also, they can contact a greater number of working professionals to improve the lead number.
- Another variable the X education can focus on is the leads whose last activity is SMS to X education. The team can make calls to this leads to be converted.
- Another strategy they can follow is contact the leads who are spending maximum time on the website.
- Based on the model analysis another strategy they can use is contacting the leads who used Olark chat. This way also we can increase the number of leads to be converted.

The X education can follow the following strategies when the company reaches its target for a quarter before the deadline.

- During this time the company can focus contacting only on high-impact lead sources like ‘Welingak Website’ and Reference.
- They can tailor strategies for Working Professionals.
- Engage with SMS and other activities effectively.
- Optimize website experience for longer visits.
- Leverage Olark Chat for real-time interactions.
- Respect 'Do Not Email' preferences.
- Monitor and adjust strategies for Landing Page Submission and Specialization_Others.

Conclusions

- The X educations can allocate more fund towards advertising in ‘Welingak Website’ as the lead conversion rate is highest for this.
- The lead conversion rate is also higher for the lead source – reference. The company can attract the learners to provide reference by giving more discounts and offers.
- They can contact the working professionals also as the lead conversion rate is high for them too, by offering placements and better packages.