Lead Scoring Case Study

Summary

X Education sells online courses to industry professionals and they have customer data, past referral, customer activity on website and other information from sources, which is classified as lead. Currently, X Education gets lot of leads but its lead conversion rate is very poor (around 30%). To be more successful and efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads' resulting to higher lead conversion reducing cost and more revenue to X Education.

During this case study, X Education expect use to do build model (Logistic Regression) to predict lead score for all customer, providing visibility of customer with higher lead score resulting higher conversion chance. The CEO has given a ballpark of the target lead conversion rate to be around 80%.

During this case study, we performed logistic regression and during the process, we faced multiple challenges. In summary, we will cover the high-level steps taken, challenges faced and learnings to overcome these challenges:

Data Cleaning and Exploratory data analysis- EDA: Post data load and sanity check, next step was to perform data cleaning. Data set has 37 columns and 9240 row with multiple null / missing values. There were lot of 'Select' field in multiple columns and in first step; 'Select' field was replaced by NaN. Columns with missing value / Null value more them 50% were dropped. There were multiple columns in data set, which were sales team generated information. We plan to drop these columns as these data were generated post conversation with customer and we would like to build model on pre-conversation data with customer. There were lot of columns with skewed data (90%, 10%) or unique data, we dropped these columns as they will not bring value to model.

Next step, we plan to impute the missing data. It was easy for columns with low missing data with mode (categorical) and median / mean (numerical) but challenge came for columns with high percentage of missing value. Post exploring option, we plan not to drop the columns but to impute data with new field 'other" as information in these columns are important for model.

EDA is core step for data modelling and outlier treatment was first activity. Soft capping of 99% was implemented for outliers. Visualization brings new information from existing data, therefore all required univariate, bivariate and correlation graphs were plotted and interesting insights were obtained (Example: Lead Source; Google is highest source category and X Education should plan better advertisement for positive lead conversion).

Data Preparation: key activity during this process was creating dummies, dividing or split data in training set (70%) and test set (30%) followed by scaling to data (Standard Scaling). During this process, we have two challenges and multiple learning:

Challenge 1: Should we drop 'other' fields from columns. We plan to drop these 'other' fields, as they will not bring any value or information to final model.

Challenge 2: While doing scaling, we were getting error (Cannot scale the test data). This was overcome by online investigation and new learning.

Model Building: During modelling process (train data set), we performed multiple activities like logistic regression model building on all features, Recursive Feature Elimination (RFE): 20 features, Manual model building, Model finalized with 13 features, Confusion matrix & Metric calculation, ROC Curve, finding Optimal Cut-off point, Precision and Recall graph and F1 Score.

During this process, Confusion matrix behaved like confusion but post upgrd session revisit and exploring internet, it worked as strength. Confusion matrix calculation was performed and outcome was positive.

Prediction on test set: Next step was to predict model and check all parameters on test set. All activity were performed and outcome was promising.

It was time to build final data set for X Education sales team with predicted conversion (0,1) and conversion probability percentage for all customers and yet again, we faced challenge: How to proceed. With sanity check activity, we append predicted train and predicted test data set followed by concatenation of 'Lead Number' to obtain final dataset for X Education sales team.

Conclusion: Finally, we had top features / predictor along with conversion probability percentage for all leads/customers. This information will help X Education sales team to increase efficiency, being effective and increase revenue by high lead conversion.

	Lead Number	Converted	Converted_Prob	Conven_Predicted	Converted_Probabilty'
0	**660737**	**0**	**0.174337**	0	17.43
1	%0728	0	0.409382		40.94
2	660727	1	0.491288		49 13
3	660719	0	0.094377	0	9.44
4	6606B1	1	0.496148		49 61
5	660680	0	0.076410	0	7.6^
6	660673	1	0.600/47		60 OF
7	660664	0	0076410	0	**7 64**
8	660624	0	0.049419	0	gg
9	660616	0	0.099967	0	10.00
10	660608		0 631839		63 18
11	**6605/0**	1	0.628110	1	62.81
12	660562		0 657438		65 7^
13	660558	0	0.139218	0	13.92
14	660553	0	0.23150*	0	23.15

TOpfeaturesIpfediCtOF

1. Lead Source: Welingak Website: 5.206469
2. Last Notable Activity: Had a Phone Conversation: 3.500256
3. What is your current occupation: Working Professional: 3.456307
4. Lead Source: Reference: 3.448623
5. Last Notable Activity: Unreachable: 1.661949
6. Last Activity: SMS Sent: 1.268065
7. Last Notable Activity: Olark Chat Conversation: 1.133033
8. Total Time Spent on Website: 1.093008
9. What is your current occupation: Unemployed: 0.936953
10. Lead Source: Olark Chat: 0.868793
11. Last Notable Activity: Modified: 0.709519
12. Lead Origin: Landing Page Submission: 0.447315
13. Specialization: Finance Management: 0.337836