LEAD SCORING CASE STUDY

Santhosh R
Anjali R S

PROBLEM STATEMENT

- An educational company sells online courses, the company uses Google, forms, websites, etc for marketing. Based on the inputs from these sources, Marketing company calls each visitor/user & make them to join the course. But the conversion rate is poor, the effort spent for conversion/phone calls are getting wasted. Need to build a machine learning model with conversion probability which will help to focus on potential customer .

APPROACH

- Data Import/Reading and Understanding the data
- Data Cleaning/EDA
- Data Preparation
- Model building
- Model evaluation
- Lead score calculation

DATA IMPORT/READING AND UNDERSTANDING THE DATA

- Dataset is given in csv format is loaded.
- The loaded dataset is read and understood.
- The data contains many null values, it has continuous and categorical variables.
- “Converted” column is the target variable.
- 9240 rows & 37 columns are present.

DATA CLEANING/EDA

- Null values and percentage of null values are calculated.
- Identified percentage of null values in column & removed column which has more than 75% of null value.
- Data imbalance has been found, column with imbalance(90, 10) has been removed.
- Outlier treatments are done.
- Remaining null values are found for Categorical variables and are replaced with mode().
- Finally obtained 9103 rows and 14 columns.

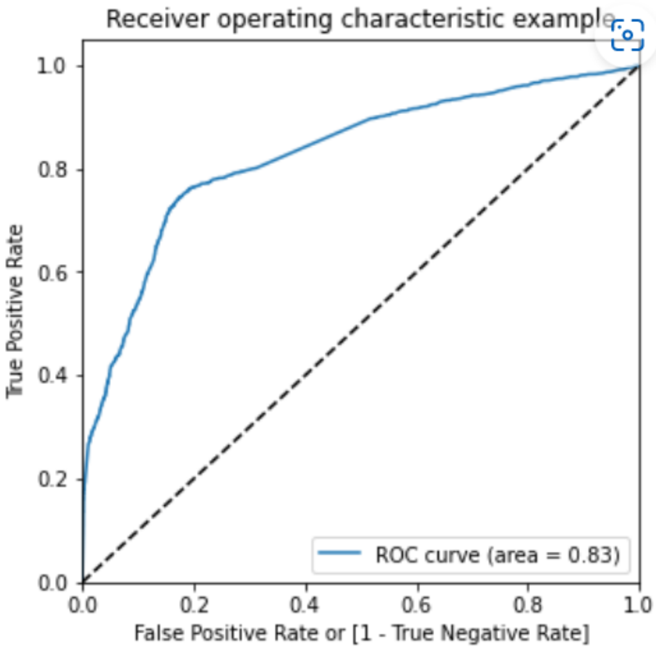
DATA PREPARATION

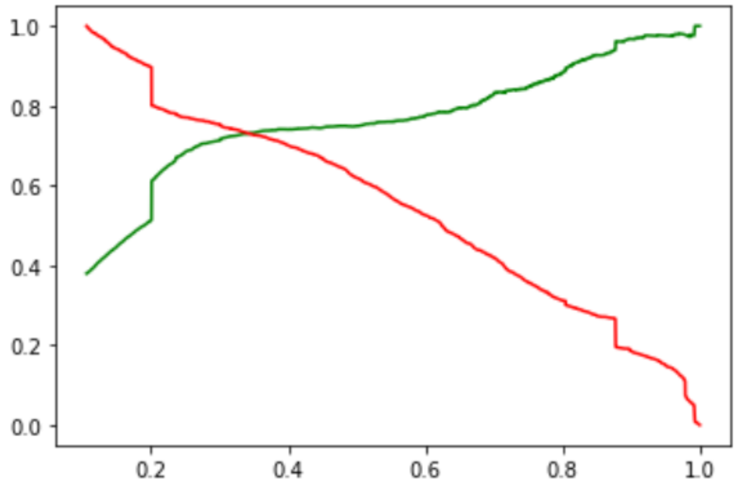
- Replaced Yes/No into 1/0.
- Dummy variables are created for categorical data (other than 0,1 – only two category).
- Removed repeated column after dummy variable creation.
- A 70%- 30% train & test data split up done.
- Scaling is performed separately for train & test data set.

MODEL BUILDING

- Variable selection using RFE is done for 18 variables.
- Checked for P-value & VIF.
- Removed columns which have high P-value(>0.05) & high VIF(>5).
- Single columns are removed at a time & the process is continued until a model with P-value & VIF are in limit.
- ROC curve value obtained is 0.83.
- Optimum cutoff point at 0.28.
- Final model’s accuracy is observed as 79%.
- Sensitivity 77%, Precision 75%, Recall 62%, Trade cut off 0.36 is obtained.

FINAL MODEL





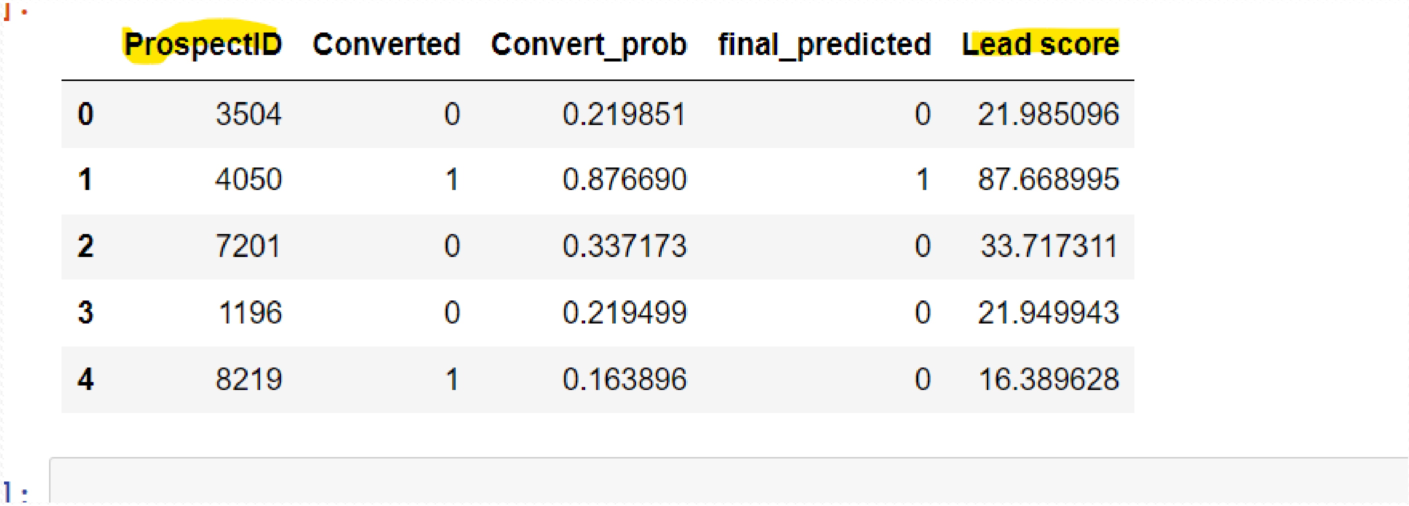
trade cutoff is 0.36

MODEL EVALUATION

- Scaling done for test data set with transform option, and no fit transform is performed.
- Conversion probability is calculated.
- Final predictive model is build based on trade cutoff calculated from train data set.
- Accuracy of the model is identified as 80% which is 1% higher than train model.
- Sensitivity 73%, Specificity 84%.

LEAD SCORE CALCULATION

- New column lead score added in the final test model "y_pred_final['Lead score'] = y_pred_final.Convert_prob.map(lambda x: x*100)"
- ProspectID with lead score has been added.



	ProspectID	Converted	Convert_prob	final_predicted	Lead score
0	3504	0	0.219851	0	21.985096
1	4050	1	0.876690	1	87.668995
2	7201	0	0.337173	0	33.717311
3	1196	0	0.219499	0	21.949943
4	8219	1	0.163896	0	16.389628

THANK YOU