Development of a Classification model with Machine Learning to predict the conversion of leads for on-line training courses.

Luis Pardina - Oct. 1st, 2022

Abstract

The dataset analyzed contains information about the leads obtained by an educational institution in India, which offers online training courses. A lead consists of many different pieces of information about a potential customer. The problem is to identify which leads get converted into actual customers.

First we performed an exploratory analysis and cleaned the data, with particular attention to handling null values; next we determined, through a study of the most relevant values and variables, the trends and behavior patterns with a high potential for lead conversion. Finally, we created and adjusted a classification model to predict lead conversion, selecting among several options Random Forest Classifier, which achieves an accuracy of 0.946.

Keywords: lead quality, supervised learning, classification model, Random Forest Classifier,

I.- INTRODUCTION

"Lead" is a concept applied in marketing & sales. In simple terms, a lead is a potential customer, an individual or organization with an interest in what you are selling (either a product or a service). Depending on the organization, the definition of the term "lead" may vary. For some companies, a "lead" is a contact already determined to be a prospective customer, whereas other companies consider a "lead" to be any sales contact. But what remains the same across definitions is that a lead will potentially become a future client. Sales and marketing teams therefore have a responsibility to convert a maximum amount of leads to maintain a good conversion rate.

The leads in e-commerce are generated when a potential customer fills up a form on the website, calls landline requesting info, initiates a chat on the website, or interacts with the company on social media. Typically that information is captured with a CRM software. CRM stands for customer relationship management. As the name suggests, CRM software is a system for managing your relationships with customers. It is used to keep track of interactions, data, and notes about customers or potential. The data is stored in a database and is accessible to multiple people within an organization.

Marketing and commercial departments typically create a lead scoring system to help them determine the quality of a lead. Knowing the quality of a lead can help a marketing team to know how likely a lead is going to convert into a sale and it can help the marketer to offer the right kind of communication to convert someone visiting the website into a paying customer.

While much is written about the adoption, usage, and failures of Customer Relationship Management (CRM), little empirical research exists to fully examine the levers to improve the conversion performance of leads in a sales funnel**[1]**. The data collected in our case study includes leads with a lot of unstructured information about those interested in online courses for an educational institution in India. We will analyze which trends and behavior patterns favor conversion, segment the potential

customers based on the probability of conversion and build an accurate prediction model**[2]**, which becomes a powerful tool to help the commercial and marketing teams to focus their efforts and maximize the sales.

II.- STATE OF ART

There are at least 3 authors[2] who have analyzed the same dataset and have developed a predictive model. The metric that drives the exercises is the accuracy: the capacity to use the existing data to make a correct prediction about the conversion or not of the lead. To summarize their process and results, they all have selected Standard Scaler to scale the variables and the Logistic Regression as a model, reaching a wide range of accuracies (0.80, 0.82 and 0.92) which is related with the very different approaches to the data cleaning and the management of the nulls.

The authors decided as a strategy to eliminate rows with null values. This reduced substantially the size of the remaining dataset to apply the model.

The Dataset is available in kaggle:

https://www.kaggle.com/code/turanmehdiyeva/lead-prediction/data

It contains 9240 instances with 37 variables and one output target (0/1). There are 6 continuous numerical variables and 31 categorical variables (either nominal, ordinal or boolean). Qualitatively, the variables can be split into two groups:

- data about the potential customer that generated the lead, such as country, city, occupation, score given by the commercial team…
- data about the interaction with the potential customer: total on-line visits, total time spent on website…

It's a balanced dataset, the target has 38,5% of positives and 61,5% of negatives.

III.- METHODOLOGY

A. Data preparation

There are no duplicate records in the data set.

Many variables, mostly of the categorical type, contain null values in percentages that can reach 30% of the total records. The general strategy has been to replace nulls with the category "unknown". The logic has been to make the most of the information contained in the variable, even if it was not complete, since deleting rows with nulls would have reduced the size of the dataset significantly. On the other hand, non-relevant categorical variables (because they have only one category) have been eliminated.

Regarding the numerical variables, we have analyzed symmetry, skew and outliers in terms of quantity and distance from the mode. Also the correlation between the numerical variables and the target variable.

B. Trends and patterns

To identify which patterns favor conversion in the leads I scan the different categories for each variable, selecting those that fulfill two conditions: to have a significant number of leads (greater than 500 leads out of 9.074) and to have a significant degree of converted leads (more that 50% when the average is 39%).

To identify which patterns favor conversion in the leads for the numeric variables, I represent them graphically in an histogram to identify the ranges where the lead conversion is higher than 50%.

C. Building a prediction model

We decided to scale the numeric variables using RobustScaler because features have many outliers and far from the mode. We converted the categorical variables to dummy variables.

It must be considered that, in a natural flow of work, the commercial teams will be interested -in the first instance- in generating the maximum number of leads possible, prioritizing quantity over quality. Therefore, what interests us is to help them in a second instance to correctly predict the quality of the lead. For this and to be able to compare with other studies carried out in the same dataset, the selected metric to evaluate the model is Accuracy: percentage of correct predictions over the total sample.

Additionally, we also present complementary metrics of the model: Precision (correct positive predictions / total positive predictions); Recall (correct positive predictions / total true positives); F1 (a mix of both precision and recall, 2*[recall * precision] / [recall + precision] and ROC/AUC curve that shows the trade-off between sensitivity/ recall (or TPR) and

False Positive Rate (1 - Specificity).

After creating train and test subsets we have tested up to 7 different models of supervised classification learning, with the standard parameter. We preselect the two that get better accuracy. The best two models regarding accuracy turn out to be Random Forest**[4]** and GradientBoostingClassifier**[5]**.

Then we applied cross-validation**[6]** in both models on the total dataset to ensure that no overfitting occurs. We have tuned the hyperparameters**[7]** to try to maximize accuracy and then we made the final selection of the model, presenting the confusion matrix**[8]**.

Once the final model is selected, we present the 20 biggest feature importances.

V.- RESULTS

A. Trends and patterns

There are 7 key indicators (shown below) of a high probability that the lead will be converted:

In the data about the potential customer:

1. When the potential customer is a working professional (Fig 1).

In the data about the interaction with the potential customer:

2. - When the lead has been originated by a form filled on-line (Fig 1).
3. - When the last activity carried out by the potential customer has been to send an SMS to X Education (Fig 1).
4. - When the lead has been tagged by the commercial team with: “Will revert after reading the email”(Fig 1).
5. - When the lead is qualified by the commercial team as: “High in relevance” or “Low in relevance” or “Might be”(Fig 1).
6. - When the total time spent by the customer on the website exceeds 800 min (Fig 2).
7. - when there are > 9 total visits to the WebSite (Fig 3).

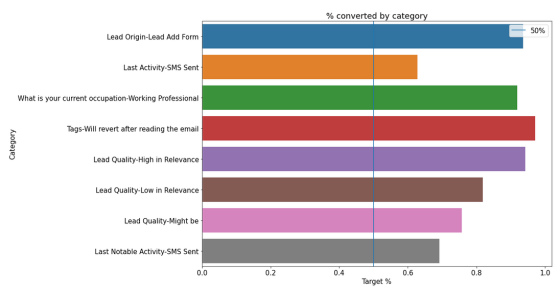

Fig.1. Categories with high conversion rate (>50%)

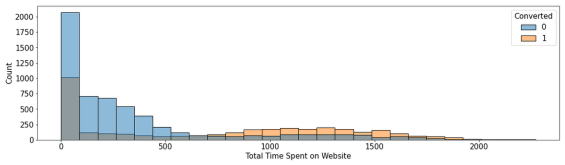

Fig.2. Conversion of leads > 50% when total time spent > 800 min

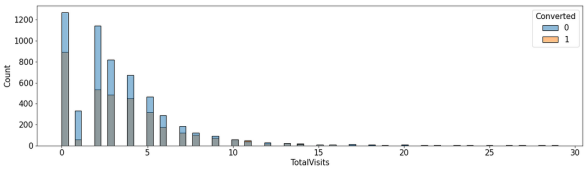

Fig.3. Conversion of leads > 50% when Total Visits > 9

B. Prediction model

The following metrics are obtained with the algorithm to make an initial test of 7 models**[9]**:

	Accuracy	Precision	Recall	F1	Roc_Auc
Logistic Regression	0.938	0.910	0.929	0.919	0.936
Random Forest	0.940	0.910	0.935	0.922	0.939
Decision Tree	0.924	0.903	0.903	0.903	0.920
K-Nearest Neighbor	0.903	0.846	0.899	0.872	0.902
AdaBoostClassifier	0.936	0.895	0.938	0.916	0.936
GradientBoostingClassifier	0.940	0.910	0.935	0.922	0.939
Balanced Bagging Classifier	0.940	0.925	0.922	0.923	0.937

Table I. Testing models metrics

The 2 models that obtain the best metrics are Random Forest and Gradient Boosting Classifier. The adjustment of the hyperparameters to optimize the accuracy results in a small improvement:

	Accuracy	Precision	Recall	F1	Roc_Auc
Random Forest	0.9401	0.9095	0.9351	0.9221	0.9392
GradientBoostingClassifier	0.9401	0.9095	0.9351	0.9221	0.9392
Random Forest Optimized	0.9431	0.9057	0.9459	0.9254	0.9436
GradientBoostingClassifier Optimized	0.9412	0.9105	0.9370	0.9235	0.9404

Table II. Metrics after hyperparameter tuning

The result of both models is very similar, although Random Forest Classifier Optimized obtains the best accuracy. Thus, this is the final selection. The accuracy achieved in the cross validation (0.946) is better than the ones of the other studies on the same dataset that were reviewed.

The best result for the precision metric corresponds to the other model, Gradient Boosting Classifier Optimized. This is not a surprise, depending on the most relevant metric selected, different models can offer the best results.

The results of the confusion matrix (predicted vs. actual) obtained with Random Forest Classifier Optimized for the accuracy metric on the test subset are shown in Fig.4:

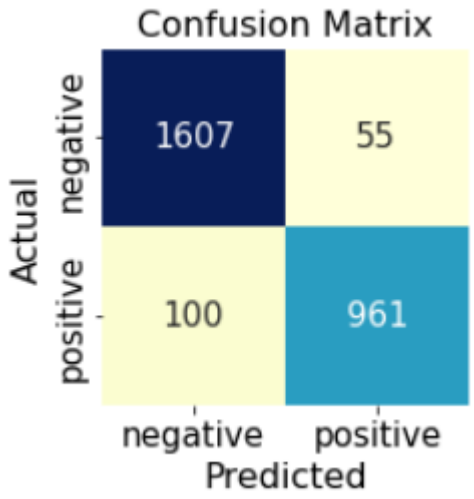

Fig.4. Confusion Matrix with selected model

We present the result of the first 20 biggest feature importances:

	Variables	Importance
91	Tags_Will revert after reading the email	0.240
2	Total Time Spent on Website	0.100
86	Tags_Ringing	0.052
125	Last Notable Activity_SMS Sent	0.045
41	Last Activity_SMS Sent	0.035
74	Tags_Closed by Horizzon	0.034
82	Tags_Lost to EINS	0.031
99	Lead Quality_High in Relevance	0.031
71	What is your current occupation_unknown	0.031
104	Lead Quality_unknown	0.031
101	Lead Quality_Might be	0.029
97	Tags_unknown	0.029
7	Lead Origin_Lead Add Form	0.027
3	Page Views Per Visit	0.017
1	TotalVisits	0.016
121	Last Notable Activity_Modified	0.016
103	Lead Quality_Worst	0.015
70	What is your current occupation_Working Profes...	0.015
100	Lead Quality_Low in Relevance	0.013
69	What is your current occupation_Unemployed	0.013

VI-. CONCLUSIONS

Obtaining the maximum possible number of leads from potential clients of a company (and subsequently registering them in the CRM) is a powerful way to increase the sales or launch new products. But the large amount of accumulated information, if it is not worked properly, can generate a dispersion of resources and unfocus the organization. This contradiction is known as "The costly conflict"**[10]** and is humorously depicted in the following cartoon (Fig. 4).

Fig.4. The Costly Conflict

Conducting telephone marketing campaigns among a company's clients is an excellent way to increase sales or launch new products. But they have the disadvantage of being expensive, they can often be annoying for customers and we also have no predictions of their outcome.

In this project we have developed a model to solve this problem in the specific case that concerns us, marketing of online courses. Firstly, through an analysis of the data, we were able to segment the potential customers based on the probability of their conversion into clients of the online courses. This will allow to establish strong guidelines to focus the marketing action. Secondly, the model achieved a prediction of the result of the lead with an accuracy greater than 0.94. This allows, among other improvements, to optimize resources, select the target leads to maximize sales and finally, it is a powerful tool to establish robust sales predictions for online courses

VII.- REFERENCES

[1] Bradford, William and Johnston, Wesley James and Bellenger, Danny, "The Impact of Sales Effort on Lead Conversion Cycle Time in a Business-to-Business Opportunity Pipeline"

(September 9, 2016). 6th International Engaged Management Scholarship Conference, 2016, Available at SSRN: https://ssrn.com/abstract=2866954 or http://dx.doi.org/10.2139/ssrn.2866954

[2] Sneha Choudhary - PG Diploma in Data Science at International Institute of Information Technology - Bengaluru, Karnataka, India: https://www.kaggle.com/code/snehac47/lead-scoring-case-study

Turan Mehdiyeva - Data Analyst - Baku, Azerbaijan: https://www.kaggle.com/code/turanmehdiyeva/lead-prediction/notebook;

Aswin Kumar - Chennai, Tamil Nadu, India: https://www.kaggle.com/code/sethuaswinkumar/lead-scoring-logistic-regression

[3] Ali Etminan Mogaji, E. and Nguyen, N.P. (2021), "Prediction of Lead Conversion With Imbalanced Data – A method based on Predictive Lead Scoring", Linköping University | Department of Computer and Information Science. Master's thesis, 30 ECTS | Datateknik | /LIU-IDA/STAT-A–21/031–SE

[4] Joaquín Amat Rodrigo. "Random Forest con Python". Octubre 2020. Available online: https://www.cienciadedatos.net/documentos/py08_random_forest_python.html.

[5] Joaquín Amat Rodrigo. "Gradient Boosting con Python". Octubre 2020. Available online: https://www.cienciadedatos.net/documentos/py09_gradient_boosting_python.html

[6] Mohammed Alhamid. "What is Cross Validation?". December 24.2020. Available online: https://towardsdatascience.com/what-is-cross validation-60c01f9d9e75

[7] Jeremy Jordan. "Hyperparameter tuning for machine learning models". November 2017. Available online: https://www.jeremyjordan.me/hyperparameter-tuning/

[8] Sarang Narkhede. "Understanding Confusion Matrix". May 9, 2018. Available online: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

[9] José Manuel Castaño. "Applying Learning Machine to predict results of a banking marketing campaign and find trends and patterns for future campaigns". September 2022. Available online: https://github.com/JoseMCastano/Banking-marketing-Classifcation-model

[10] Rajat Arora. "What is a lead?". August 2022. Available online: https://www.leadsquared.com/what-is-a-lead/