Data Science Final Project

Lead Quality Prediction Model

by Luis Pardina, July 28th 2022.

1. Presentation of the chosen data set:

What is a lead?

It is a concept applied in marketing & sales. In simple terms, a **lead** is an individual or organization with an interest in what you are selling (a product, a service, ...). The interest is expressed by sharing contact information, like an email ID, a phone number, or even a social media handle. So, a lead is generated when someone fills up a form on your website, calls your company landline requesting info, initiates a chat on your website, or interacts with you on social media.

A lead is a potential customer who may convert (became an actual customer)... or not.

Marketing 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 your website or wandering through your store into a paying customer.

Content of the data set

The data set contains leads collected by *X Education* which sells online courses to industry professionals.

The information is about the source of each lead, characteristics of the lead (such as time spent in the website or number of visits to its pages), and data about the characteristics of the potential customers collected when these people fill up a form providing info about their quest.

It also explains if each lead has been converted into an actual customer or not. **69%** of the total number of leads in the data set were converted.

2. General characteristics:

The Dataset is available in kaggle:

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

The Dataset contains **9240** records with **37** variables.

There are 6 continuous numerical variables, and 31 categorical variables (either nominal, ordinal or boolean).

3. Description of the content of the variables:

Variables	Description
Prospect ID	A unique ID to identify the customer.
Lead Number	A lead number assigned to each lead procured.
Lead Origin	The origin identifier with which the customer was identified to be a lead. Includes API, Landing Page Submission, etc.
Lead Source	The source of the lead. Includes Google, Organic Search, Olark Chat, etc.
Do Not Email	An indicator variable selected by the customer wherein they select whether or not they want to be emailed about the course or not.
Do Not Call	An indicator variable selected by the customer wherein they select whether or not they want to be called about the course or not.
Converted	The target variable. Indicates whether a lead has been successfully converted or not.
TotalVisits	The total number of visits made by the customer on the website.
Total Time Spent on Website	The total time spent by the customer on the website.
Page Views Per Visit	Average number of pages on the website viewed during the visits.
Last Activity	Last activity performed by the customer. Includes Email Opened, Olark Chat Conversation, etc.

Country	The country of the customer.
Specialization	The industry domain in which the customer worked before. Includes the level 'Select Specialization' which means the customer had not selected this option while filling the form.
How did you hear about X Education	The source from which the customer heard about X Education.
What is your current occupation	Indicates whether the customer is a student, unemployed or employed.
What matters most to you in choosing this course	An option selected by the customer indicating what is their main motto behind doing this course.
Search	Indicating whether the customer had seen the ad in any of the listed items.
Magazine	
Newspaper Article	
X Education Forums	
Newspaper	
Digital Advertisement	
Through Recommendations	Indicates whether the customer came in through recommendations.
Receive More Updates About Our Courses	Indicates whether the customer chose to receive more updates about the courses.
Tags	Tags assigned to customers indicating the current status of the lead.
Lead Quality	Indicates the quality of lead based on the data and intuition of the marketer / sales who has been assigned to the lead.
Update me on Supply Chain Content	Indicates whether the customer wants updates on the Supply Chain Content.
Get updates on DM Content	Indicates whether the customer wants updates on the DM Content.

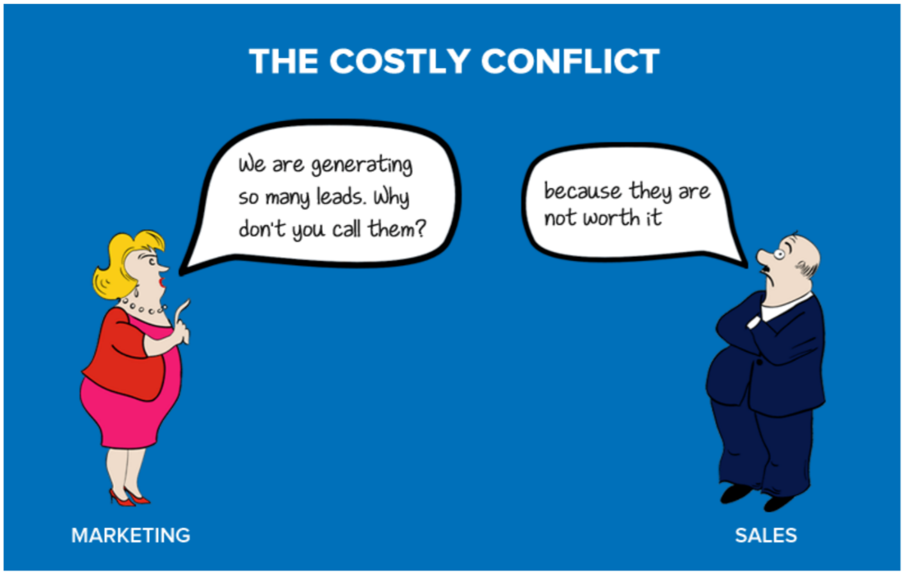
Lead Profile	A lead level assigned to each customer based on their profile.
City	The city of the customer.
Asymmetrique Activity Index	An index and score assigned to each customer based on their activity and their profile
Asymmetrique Profile Index	
Asymmetrique Activity Score	
Asymmetrique Profile Score	
I agree to pay the amount through cheque	Indicates whether the customer has agreed to pay the amount through cheque or not.
a free copy of Mastering The Interview	Indicates whether the customer wants a free copy of 'Mastering the Interview' or not.
Last Notable Activity	The last notable activity performed by the student.

4. Objectives of the project:

Build a **classification model** to predict the most promising leads, i.e. the leads that are most likely **to convert** into paying customers.

Find **trends and patterns** from the information provided in the leads.

Maximize the **accuracy** and **precision** of the model, focusing in **optimizing the effort** that has to be made by sales and marketing teams with the leads, as expressed in “The costly conflict” below:


THE COSTLY CONFLICT
We are generating so many leads. Why don't you call them?
because they are not worth it
MARKETING
SALES

State-of-art: As a reference other projects working with the same dataset applying classification models achieve accuracies > 80% and precisions around 80%