Machine Learning Approach to Predict Sales Based on Lead's Lifecycle.

Daniel Palomino

February 25, 2021

Abstract

We present a brief report for the challenge proposed by Tranzact company. For a deeper review of analysis and results, please refer to the notebooks located in the repository of the present project [1].

1 Analysis

As part of the present challenge, two datasets were provided: Call Facts and Lead Facts. A brief description of each column in datasets are presented in Appendix A.
Using Lead Facts dataset, although there are a total of 27751 records, only 3.9% are converted on sales.

1.1 Insights

With respect to Call Facts dataset:

1. Calls are uniform across days of the week and will not be useful in this analysis.
2. Days of week do not seem relevant to predict a sale.
3. Although fewer calls made in the second third of the year, a better conversion rate was achieved.

With respect to Lead Facts dataset:

1. The source system with higher conversion rate is CUY-AMY. However, tztLeads provides a higher quantity of leads.
2. IB contacts are more relevant than OB.
3. Rate conversion in people under 20 is higher than other groups.
4. Leads with delivery method Voice and lead type Shared can achive a sale easier than other groups.
5. Calls are uniform accross days of weekdays.
6. Day of the week does not seem relevant to predict a sale.

[1] https://github.com/dpalominop/SalesPredictor

2 Modelling

For the present work, it has been proposed an approach based on a Machine Learning algorithm using LSTM cells as core. This kind of recurrent network was chosen because the the lead's lifecycle is composed by several previous states and future ones are dependent on previous steps.

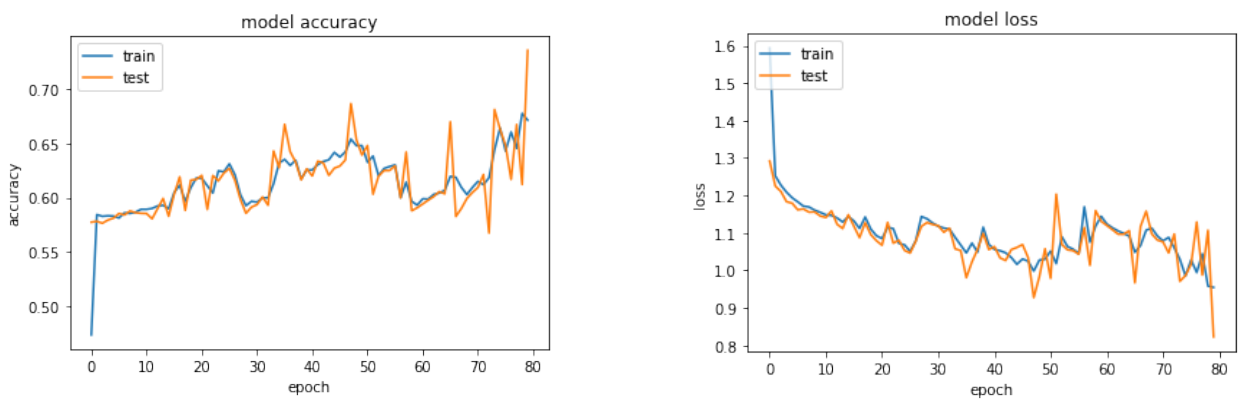
2.1 Model

Since this is a proof of concept, it is more convient to use a simple architecture as shown in Table 1:

Layer (type)	Output Shape	Param #
LSTM	(None, 22, 31)	1922-1930
LSTM	(None, 22, 64)	1930-1936
LSTM	(None, 64)	1946
Dropout	(None, 64)	1947
Dense	(None, 8)	1954
Activation	(None, 8)	1958
Total params: 65,932		
Trainable params: 65,932		
Non-trainable params: 0		

Table 1: Architecture used to build the model presented in this work.

After training this model for 80 epochs, the results plotted in Fig. 1 are gotten.



(a) (b)

Figure 1: (a) Accuracy for training and validation datasets. (b) Loss for training and validation datasets.

2.2 Iterpretation

In order to get a better understanding, a Classification Report is shown in Table 2.

3 Summary and Conclusions

From previous analysis we get some customer segments are not worked properly. For example, people under 20 is the group with less members. However, it has the highest conversion rate at all as shown in Fig. 2.

Similar cases are found when we analyse the attributes leadType and deliveryMethod. The groups with less people are the group with the highest conversion rate.

	precision	recall	f1-score	support
Agent Call Back	0.00	0.00	0.00	109
Do Not Call	0.00	0.00	0.00	40
No Sale	0.74	0.98	0.84	2040
Non Workable	0.76	0.64	0.70	663
Sale	0.87	0.37	0.52	217
Still Workable	0.85	0.96	0.90	1940
System disposition	0.00	0.00	0.00	34
Transfer Call	0.00	0.00	0.00	508
accuracy			0.79	5551
macro avg	0.40	0.37	0.37	5551
weighted avg	0.69	0.79	0.73	5551

Table 2: Classification Report using precision, recall and f1 score.

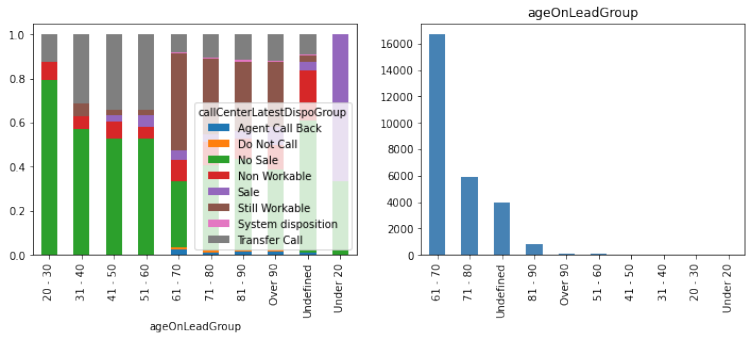


Figure 2: People grouped by age.

On the other hand, the model proposed in this work can be considered successful because although it has an f1 score of 0.52 on Sale objective, it achieved 0.87 in precision metric, which means almost all leads that were predicted as sales were correct.

Finally, this work has proved that it is possible to increase sales using the previous and actual stated of the lead's life cycle. It will be helpful because the company will be able to filter leads based on the most probable future sales.

Appendix A Datasets

A.1 Call Facts

Last telephone Contact:

- **dialerLeadId**: id of original Lead (numeric).
- **dimAgentId**: id of who contacted the client (numeric).
- **callDate**: begining of contact (categorical: "1/4/2018 5:29:53 PM").
- **callEndDate**: ending of contact (categorical: "1/4/2018 5:30:11 PM").
- **callType**: type of call (categorical: "Inbound", "Outbound", "Manual Dial", "Outbound Agent Callback", "Voicemail").
- **callTypeId**: id of callType (numeric).
- **dispoGroup**: result of actual call (categorical: "Agent Call Back", "Do Not Call", "No Sale", "Non Workable", "Sale", "Still Workable", "System disposition", "Transfer Call").

- **dispoDescription**: description of dispoGroup (categorical: "Answering Machine", "Busye", ..., "Sale").
- **callAttemptNumber**: number of attempts (numeric).
- **talkTimeSeconds**: time of call in seconds (numeric).
- **handleTimeSeconds**: handled time in seconds (numeric).
- **outboundCalls**: number of outbounding calls (numeric).
- **outboundCallbacks**: number of outbounding callbacks (numeric).
- **inboundCalls**: number of inbounding calls (numeric).
- **manualCalls**: number of manual calls (numeric).
- **inboundCallsHandled**: number of inbounding calls (numeric).
- **callSkill**: unknown data (numeric).

Other Attributes:

- Other attributes were not included in this analysis because they were meaningless or repetitive data.

A.2 Lead Facts

Client Attributes:

- **ageOnLeadGroup** : age (categorical: "Under 20", "20 - 30", "31 - 40", "41 - 50", "51 - 60", "61 - 70", "71 - 80" , "81 - 90", "Over 90", "Undefined")
- **gender**: gender (categorical: "F", "M", "U")

Lead Attributes:

- **dialerLeadId**: id of Lead (categorical: "1/1/2018 12:47:18 AM").
- **sourceSystem**: id of system which provided the lead (categorical: "CUY-AMS", "PolicyFile", "tztLeads").
- **leadDate**: date of creation of lead (numeric).
- **afid**: unknown data (numeric). campaign id: id of campaign (numeric).
- **originalContactType**: type of contact (categorical: "IB", "OB", "Undefined").
- **leadTypeId**: id of type of lead (numeric).
- **leadZip**: zip of lead (numeric).
- **leadCost**: cost of lead (numeric).
- **callCenterLeadCreatedReason**: lead created reason (categorical: "Clone:OnCall", "Draft", ..., "Import").
- **callCenterLeadCloneSourceId**: lead clone source id (numeric).
- **leadsReceived**: number of leads received associated to client (numeric).
- **leadsAccepted**: number of accepted leads associated to client (numeric).
- **leadsRejected**: number of rejected leads associated to client (numeric).
- **callCenterLeadsTotal**: number of total leads associated to client (numeric).

- **callCenterLeadEverContacted**: number of total leads ever contacted associated to client (numeric).
- **callCenterLeadEverSold**: number of total leads ever sold associated to client (numeric).
- **currentSkill**: unknown data (numeric).
- **originalSkill**: unknown data (numeric).
- **expirationInDays**: expiration of lead (numeric).
- **maxAttempts**: maximum attempts to contact (numeric).
- **dailyMaxAttempts**: maximum attempts by day (numeric).
- **leadType**: type of lead (categorical: "Primary", "Recycle", "Shared").
- **deliveryMethod**: type of communication (categorical: "Data", "Voice").
- **everCallback**: unknown data (numeric).

Other Attributes:

- Other attributes were not included in this analysis because they were meaningless or repetitive data.