ESILV
ENGINEERING SCHOOL
DE VINCI PARIS

Advanced Machine Learning

Project Report

Students :
KUOCH Jacky
NICOLAS Kevin
ALIZADEH Charly
LIMNAVONG Thomas

DIA 1

Subject

On the course module folder you will find a dataset listing purchase transactions. Adapt to the capacity of your computer (work on sampling possible). The objective is to model the income (transactionRevenue) generated per person (fullVisitorId).

The challenge is not to have the best performance but on the contrary to have the most thoughtful and structured approach to address the problem The data was uploaded on October 26 and the assignment is due on Monday December 20 at 11:59 p.m. at the latest The project is to be carried out in teams of 4 maximum.

Your report will contain :

- **Your analysis of the problem, the data, the description of your approach to solve it, the algorithms tested, their results in terms of performance and the importance of the variables (At least 3 algorithmic approaches required).**
- **Models you tried and how you tuned them. Your best prediction evaluation.**
- **Your assessment on the best way to solve the problem and the new avenues that you could test with more time.**
- **Format: Word or PowerPoint version, a PDF if other writing format.**
- **4 or 5 pages minimum, as explicit as possible: as if you were returning it to your business client within the company.**

1 Problem Analysis

The problem we are facing is a regression one, meaning that we need to predict a continuous target. The features are both continuous and categorical, for the categorical features we'll need to convert them to numerical features by using dummy/indicator variables.

2 Data Analysis

Our dataset contains 55 columns and 903.653 rows. As we are studying purchase transactions, we have information like the date and time of the purchase, the location of it but also the browser or operation system on which the transaction took place.

2.1 Location

As consumption trends may vary a lot according to the regions of the world but also meet depending on communities' affinities, it is important to consider the various locations that our data offers us when trying to study the behaviours behind purchase transactions.

Let's have a look at the continents, subcontinents, countries and cities of our transactions. We will start by studying on a global scale and observe the number of transactions per continents:

Continent	**Count**
Americas	450.377
Asia	223.698
Europe	198.311
Oceania	15.054
Africa	14.745
(not set)	1468

We easily see that our dataset offers a majority of American transactions as Americas' transactions numbers represent almost the double of transactions of the second continent most represented: Asia. Europe comes in third with a still relevant number of transactions before Oceania and Africa with numbers much lower than the three first continents.

2.2 Digital tools

We also have many information about the devices and tools used by the user to realize their transaction including the web browser, the operating system and if they are using their phone or their desktop.

Browser	Count
Chrome	620.364
Safari	182.245
Firefox	37.069
Internet Explorer	19.375
Edge	10.205
...	..

Device	Count
Desktop	664.479
Mobile	208.725
Tablet	30.449

Operating System	Count
Windows	350.072
Macintosh	253.938
Android	123.892
iOS	107.665
Linux	35.034
Chrome OS	26.337
(not set)	4.695
...	...

Source	Count
Google	400.788
Youtube	212.602
(direct)	143.028
mall.googleplex.com	66.416
Partners	16.411
...	...

Looking at this could explain to us which tools offer the best User Experiences and have more chance to convert a visit into a purchase. Nowadays, it is really difficult to guide a consumer through all the steps starting from the user reaching our platform, then having a look at our products, convince him to buy the product and then the final step, the payment. Many means are implemented so that the customer experience is as optimized as possible.

Therefore, knowing which browser, which operating system or the source where the customers come from is a a precious piece of information so that companies can invest in the fields where their customers are most likely to make a purchase.

3 Models

3.1 Cleaning the data

Before applying our different algorithms, we had some cleaning to do on our data. First, we noticed that our target was missing a lot of values. We considered that these rows were visits that were not converted into transactions, therefore we set a 0 values for all the rows missing a target value.

We also made this assumption for the following columns:

- isTrueDirect
- page
- adNetworkType
- newVisits
- bounces
- pageviews

Single-value columns are useless for machine learning therefore we dropped them.

We also dropped the following columns because we judged that they didn't contain pertinent informations, or because too much observations had them.

- metro
- region
- city
- sessionId
- visitId
- adContent
- isVideoAd
- slot
- gclId
- keyword

- campaign

Then we converted the categorical columns into numerical's. Those columns are:

- channelGrouping
- source
- medium
- continent
- subContinent
- country
- browser
- operatingSystem
- deviceCategory
- networkDomain
- referralPath
- isTrueDirect
- adNetworkType

We also converted the date column into DateTime python objects and extracted the day of the week and month, then we removed the date column.

3.2 Feature Engineering

Before grouping the dataset by the fullVisitorId we added one feature called timeSpanSinceFirstVisit which compute the time difference between the first time a user visited the shop and the current visit of the shop.

When grouping the dataset by fullVisitorId we had to make choices for the aggregation of the different columns. You can find those aggregation choices in the table 1

Feature	Aggregation function(s)
channelGrouping	max
visitNumber	max
visitStartTime	max
source	last
medium	last
isTrueDirect	last
referralPath	last
hits	sum, 'mean', 'min', 'max', 'median'
pageviews	sum, 'mean', 'min', 'max', 'median'
bounces	sum, 'mean'
newVisits	max
transactionRevenue	sum
continent	last
subContinent	last
country	last
networkDomain	last
browser	last
operatingSystem	last
isMobile	last
deviceCategory	last
page	last
adNetworkType	last
dayofweek	last
month	last
timeSpanSinceFirstVisit	last

Table 1: Per column aggregation functions

3.3 Split between train, val and test

We split the dataset between train, val and test set using 80%, 10% and 10% of the data respectively. We used the function $\text{train}_t est_s plit from sklearn to have a similar distributi$

3.4 Models

3.4.1 Random Forest

First we trained a binary classifier to predict wether an observations have a null or non null target. To do so we used a random forest classifier. You can find the results bellow.

577078	0
1	8177

Table 2: Confusion matrix for the train set

14370	57
134	71

Table 3: Confusion matrix for the validation set

Accuracy	F1
0.99	0.99

Table 4: Train scores

Accuracy	F1
0.99	0.43

Table 5: Val scores

3.4.2 Feed Forward Neural Network

Then we trained a neural network on the undersampled. You can find the evolution of the loss on fig. 1 and a visualization of the model performance on

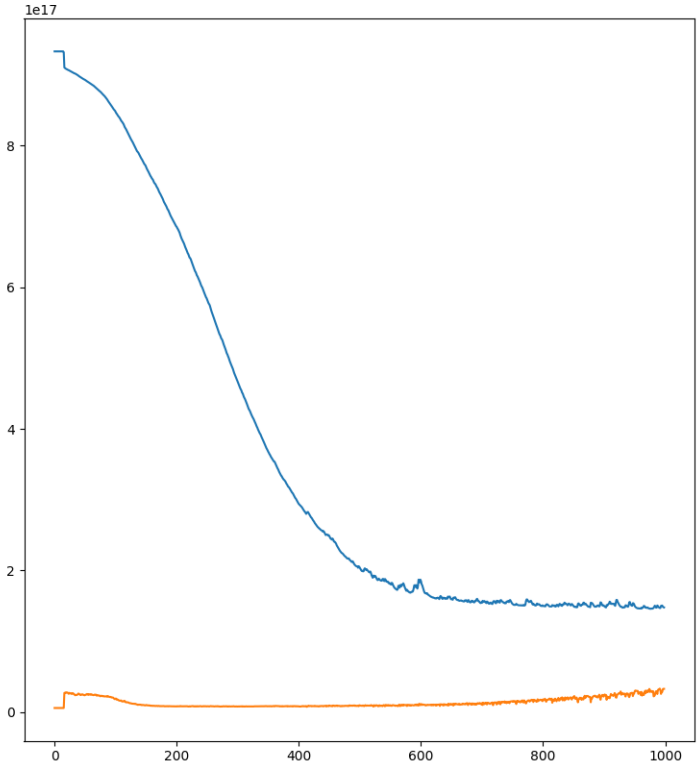

Figure 1: Evolution of the loss, in blue the training loss and in orange the validation loss

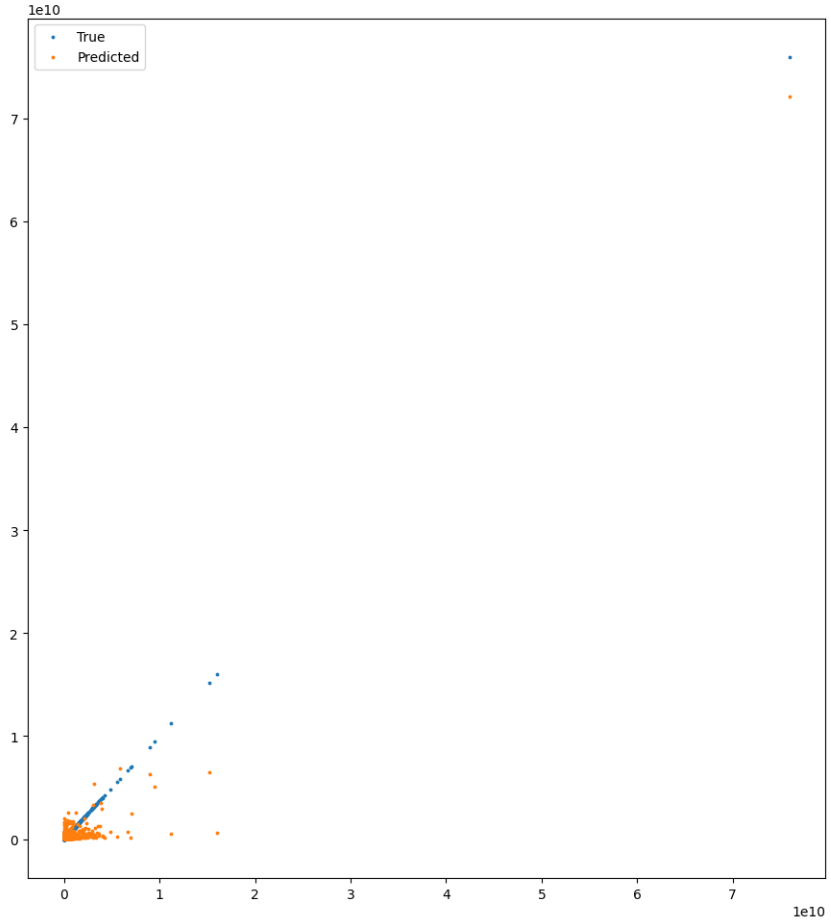

Figure 2: Train target visualization

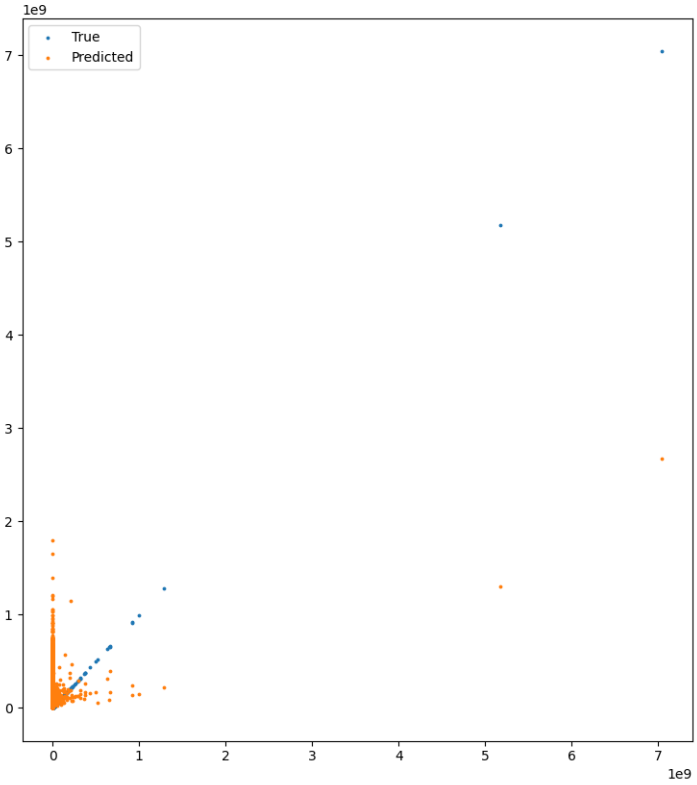

Figure 3: Val target visualization

As you can see the model performs very badly, we tried to improve it but could achieve any improvement.

3.5 Final model

To predict the target we used both models. First we predict which observations will yield to a non null target using the random forest classifier, then we use the neural network to make prediction on those observations. You can find the R2 scores in

Train	Val	Test
0.81	0.41	0.13

Table 6: R2 scores