Training Test Data Overflow 2020-2021

MARWANE ADALA (SUPCOM)

I-First Part:

1) What makes you think that you can become a great data scientist?

Very comfortable with numbers and statistics, I like giving them meaning and seeing them impact reality. Data science was therefore an ideal discipline for me. To achieve this, I turned to devote my time to master the various corresponding aspects.
I had the opportunity to carry out a few missions in data science as part of my summer internships and academic projects and as part of my online work as a teacher on the French platform www.supadom.fr. These short experiences allowed me to discover in a short time different markets, different types of companies, and different analysis programs. Each had its objectives, its challenges, but all needed a compilation and analysis of data provided by the Internet and social networks.
Finally, I should mention that manipulating data is essential to make progress nowadays in any field and Data Science is a powerful tool to get in-depth insights which is really marvelous.

2) What is the difference between Supervised Learning and Unsupervised Learning?

In the field of Machine Learning, there are two main types of learning: supervised and unsupervised. The main difference between the two types is that supervised learning is done on the basis of a truth. In other words, we have prior knowledge of what the output values of our samples should be. Therefore, the goal of supervised learning is to learn a function that, from a sample of data and desired outcomes, best approximates the relationship between observable input and output in the data. In contrast, unsupervised learning does not have labeled outcomes. Its objective is therefore to deduce the natural structure present in a set of data points.

To sum up:
Supervised: All data is tagged and algorithms learn to predict the outcome of the input data.
Unsupervised: All data is not labeled and algorithms learn the inherent structure from the input data.

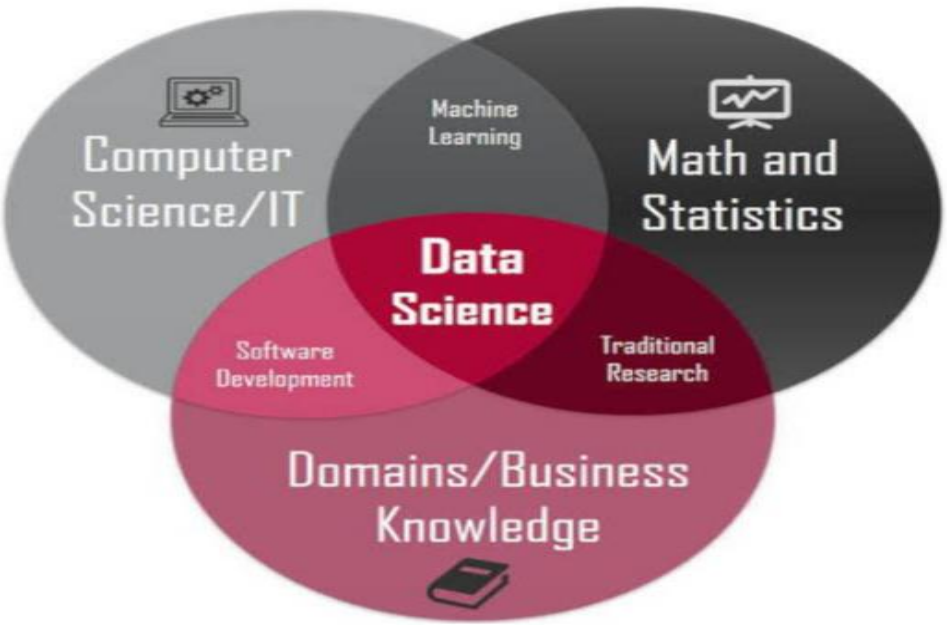
3) Differentiate between Data Science, Machine Learning and AI.

*Data Science (this is big word that involves many things)

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and ideas from many structural and unstructured data.

It uses techniques and theories drawn from many fields in the context of mathematics, statistics, computer science, theory and information technology.

Among them: probabilistic models, machine learning, statistical learning, computer programming, data engineering, pattern recognition, data visualization, uncertainty modeling, storage data, geo-visualization, data compression and high performance computing. Methods that adapt to big data are of particular interest in data science, although the discipline is not generally considered to be limited to such data.



Data Scientist's job is to help business issues to be addressed more effectively using data insights. Most of the time, business problems in this day and age can be solved with simple descriptive or curious information, as they were blank of data in the past. Otherwise, we try to optimize them further with predictive models and begin to build more normative ones to conduct business more efficiently.

*Machine Learning:

This is a science in which the machine / computer is programmed to learn. “Learning” is where people start to lose it. Learning can come from a static data set or it can be continuous learning. For example, consider a child who started learning to walk seeing his parents walk against him who evolved as a runner and learned how to improve his running every day.

Machine learning is a form of artificial intelligence (AI) that allows a system to learn from data, not through programming. However, machine learning is not a simple process. As the algorithms ingest the training data, it becomes possible to create more accurate models based on this data. A machine learning model is the output generated when you train your machine learning algorithm with data. After training, when you provide input data to a model, you receive an output result. For example, a predictive algorithm creates a predictive model. Then, when you provide data to the predictive model, you receive a forecast that is determined by the data that was used to train the model.

*AI

Artificial intelligence is the field in which the machine is constantly learning from the events that you feed it, while machine learning is a broader term. The people who create a unique prediction model for weather forecast based on data from previous years are not building an artificially intelligent system but a machine learning model, while the people who create a prediction model for weather forecast which improving every day with the arrival of new data are building an artificial intelligence model that is also machine learning.

4) Let's suppose you are a data scientist at a real estate agency. You have all the data concerning house pricing. How are you going to handle your data and what will you make of it?

For this case, I will follow the next steps with dealing with this data:

1. Discover and assess data

After collecting the data, it is important to discover each dataset. This step is about getting to know the data and understanding what has to be done before the data becomes useful in a particular context. Discovery is a big task, but it is made easily with the data preparation platforms that we have nowadays offering visualization tools which help users profile and browse their data.

2. Cleanse and validate data

Cleaning up the data is traditionally the most time consuming part of the data preparation process, but it's crucial for removing faulty data and filling in gaps. Important tasks here include:

- Removing extraneous data and outliers.
- Filling in missing values.
- Conforming data to a standardized pattern.
- Masking private or sensitive data entries.

Once data has been cleansed, it must be validated by testing for errors in the data preparation process up to this point. Often times, an error in the system will become apparent during this step and will need to be resolved before moving forward.

Examples how to apply steps 1 and 2 above

Here, I should mention a step-by-step process of how to build house price predictions.
For this I suggest using the next technologies python libraries 1) Numpy 2)pandas 3) Matplotlib 4) Sklearn

Lets start with data, this dataset can contain some basic information of houses like area type, location, size, price, etc...
First step is we remove unnecessary data from this data set which no affect on price estimate. So after removing that data, our dataset is smarter and easy to interpret.
Now we start with Data cleaning process, we check is there any null values in dataset or not?
We can also split the size features data with space and make new features.
"Remove outliers": We should remove outliers using for example mean and one standard deviation.

3.Transform, enrich and store data

Transforming data is the process of updating the format or value entries in order to reach a well-defined outcome, or to make the data more easily understood by a wider audience. Enriching data refers to adding and connecting data with other related information to provide deeper insights.

Once prepared, the data can be stored or channeled into a third party application—such as a business intelligence tool—clearing the way for processing and analysis to take place.

II- Second Part:

Given such a dataset of grocery sales for two years. 'Member_number' = id of customer,

'Date' = date of purchase,

'itemDescription'= Description of product purchased. Explain what will you do with this dataset as a data scientist.

As a data Scientist, I will follow the steps mentioned in last question of the First Part of this test. Here, we intend to reach the next objectives:

- Product Recommendations
- Assortment Optimization
- Pricing
- Personalized marketing

To go through these objectives, we need (after applying the steps mentioned above) to build a model that tells us:

-which items are most purchased in each cycle so that the store can supply the quantity in need,

-which items should be put together on grocery rows,

-how to plan the pricing of the different items and make a personalized marketing for each customer.

In this question, we can also forecast product sales based on the items Description and date of purchase mainly. (Member ID can be helpful if we want to get insights about the improvements done concerning the sales and the customer satisfaction: we can for example make a range of Ids for the first year and another range for the second year and tell the difference between the two years).

We should provide the grocery store with the possibility to please customers by having just enough of the right products at the right time. For this particular problem, we can analyse the data as a supervised learning problem. In order to forecasts the sales we can compare different regression models like Linear Regression, Decision Tree, ExtraTreeRegressor, Gradient Boosting, Random Forest and XgBoost. Further to optimize the results we can use multilayer perception (MLP: a class of feed forward artificial neural network) and LightGBM (gradient boosting framework that uses tree based learning algorithms).

Using libraries such as pandas, matplotlib, numpy, sklearn,.. etc, we can get can powerful results and visualizations:

For example, follow this lines of python code:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
from sklearn import metrics
purchase=pd.read_csv('...../Grocery/purchase.csv')
purchase.shape
```

```
purchase.sort_index(inplace=True)
purchase.head()
```

```
user_item_counts =
purchase.groupby("Member_number").apply(item_counts).unstack(fill_valu
e=0)
#we can build the function item_counts that calculates to bought items
for each customer based on Member_number.
```

The customer who bought the most items overall in her lifetime

```
user_item_total = user_item_counts.sum(axis=1)
```

For each item, the customer who bought that product the most

```
max_item = user_item_counts.apply(lambda s: pd.Series([s.argmax(),
s.max()], index=["max_user", "max_count"]))
max_item = max_item.transpose()
```

Cluster items based on user co-purchase history. That is, create clusters of products that have the highest probability of being bought together.

And so…. We can also work with Date column to determine periods were there are lots of purchase…

	Member_number	Date	itemDescription
0	1808	21-07-2015	tropical fruit
1	2552	05-01-2015	whole milk
2	2300	19-09-2015	pip fruit
3	1187	12-12-2015	other vegetables
4	3037	01-02-2015	whole milk
5	4941	14-02-2015	rolls/buns
6	4501	08-05-2015	other vegetables
7	3803	23-12-2015	pot plants
8	2762	20-03-2015	whole milk
9	4119	12-02-2015	tropical fruit
10	1340	24-02-2015	citrus fruit
11	2193	14-04-2015	beef
12	1997	21-07-2015	frankfurter
13	4546	03-09-2015	chicken
14	4736	21-07-2015	butter
15	1959	30-03-2015	fruit/vegetable juice

16	1974	03-05-2015	packaged fruit/vegetables
17	2421	02-09-2015	chocolate
18	1513	03-08-2015	specialty bar
19	1905	07-07-2015	other vegetables
20	2810	08-09-2015	butter milk
21	2867	12-11-2015	whole milk
22	3962	18-09-2015	tropical fruit
23	1088	30-11-2015	tropical fruit
24	4976	17-07-2015	bottled water
25	4056	12-06-2015	yogurt
26	3611	13-02-2015	sausage
27	1420	14-01-2015	other vegetables
28	4286	08-03-2015	brown bread
29	4918	27-01-2015	yogurt
...	...	...	...
38735	4290	26-03-2014	instant coffee
38736	1818	08-12-2014	beverages
38737	1176	04-08-2014	bottled water
38738	4879	19-09-2014	zwieback
38739	1574	17-06-2014	pastry
38740	1045	10-04-2014	Instant food products
38741	1168	31-10-2014	long life bakery product
38742	4648	03-10-2014	whipped/sour cream
38743	1931	04-03-2014	salt
38744	2868	04-03-2014	potato products
38745	3082	22-07-2014	whole milk
38746	2935	21-04-2014	frozen vegetables
38747	2639	08-06-2014	fruit/vegetable juice
38748	2789	12-07-2014	sugar
38749	4153	14-02-2014	chocolate
38750	3761	12-02-2014	sugar
38751	4444	25-04-2014	pastry
38752	2824	05-09-2014	cling film/bags
38753	1146	12-07-2014	waffles
38754	4796	02-03-2014	Instant food products
38755	4586	26-09-2014	bottled water
38756	1987	29-10-2014	fruit/vegetable juice
38757	4376	07-12-2014	rolls/buns
38758	2511	18-06-2014	long life bakery product
38759	3364	06-05-2014	oil
38760	4471	08-10-2014	sliced cheese
38761	2022	23-02-2014	candy
38762	1097	16-04-2014	cake bar
38763	1510	03-12-2014	fruit/vegetable juice
38764	1521	26-12-2014	cat food

38765 rows × 3 columns