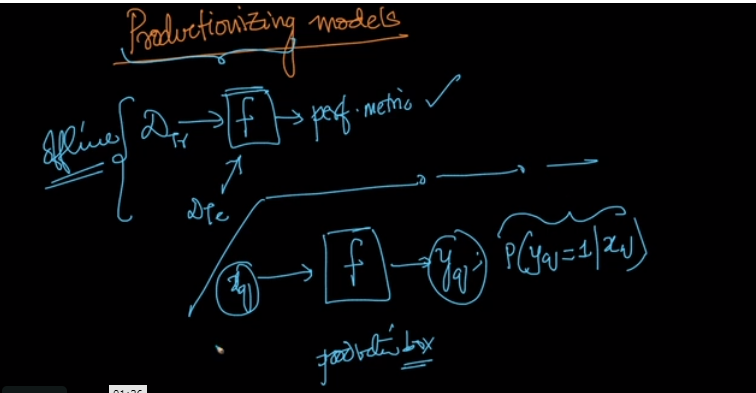
Now it comes to productionizing model i.e. when everything is done for preparing model and our performance is also acceptable then we want our model to take real world inputs and this

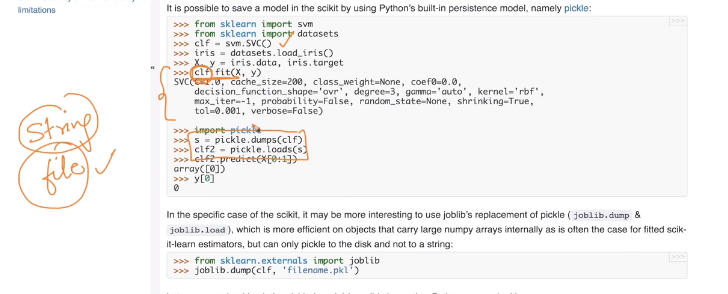
Will be done on production box where in some case it will take input and return Yi, in some cases P(Y = 0|X).



So there are two ways to do it.

First one is in scikit learn itself called persistence model i.e. your trained model will be stored in memory and whenever you want to run your model just load it to RAM and then

You give it Xq It will return Yq.



There are two formats in which we can store our model i.e. string or file.

Recommended to store in file format as large models can’t be stored in string format.

To store we use pickle or another better way is to use joblib.



And the model is stored in file format with extension .pkl

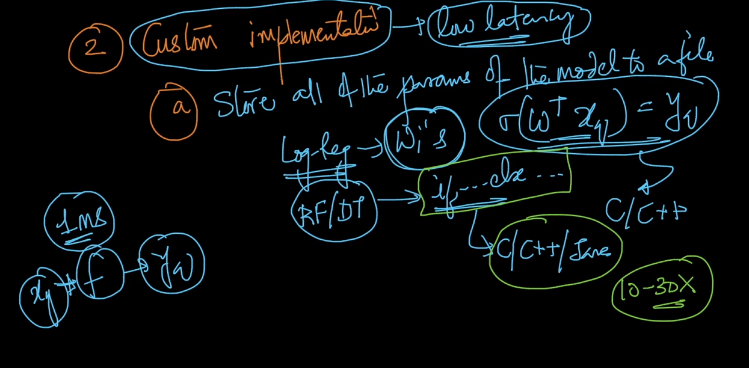
Now another way of productionizing model is custom implementation and what it means is

For any model store all its parameters as we want to store instead of storing in pre-implemented approach.

For example for LR store weights , For DT store it in if-else format

And why it is needed is because we may be having low latency requirements and python takes time.

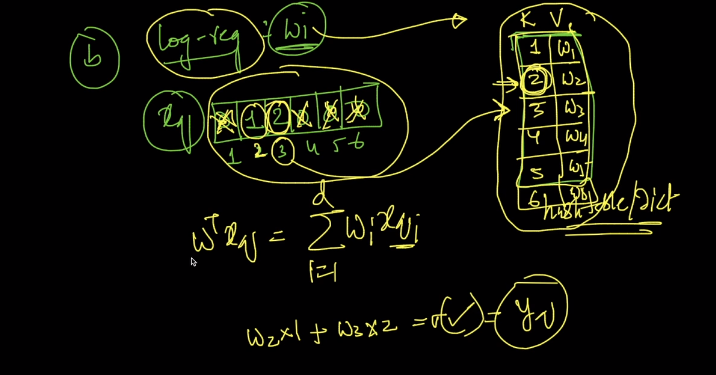
SO instead of storing it in python using C/C++/Java can boost up speed.



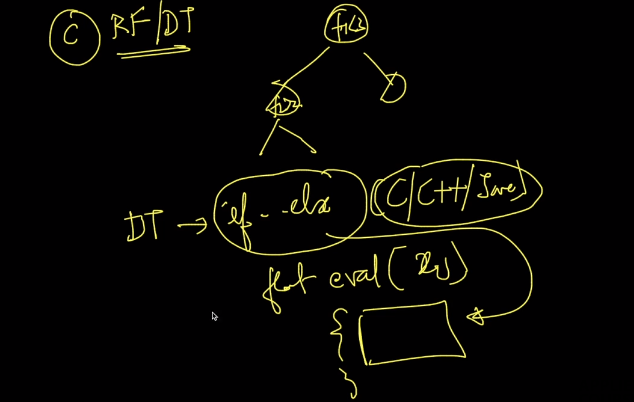
Now lets take an example:

Suppose we have our LR and we store our weights in Hashtable/Dictionary in Java/Python.

SO when we get our Xq we just multiply its value by its corresponding weights and then calculate sigmoid and determine Yi.



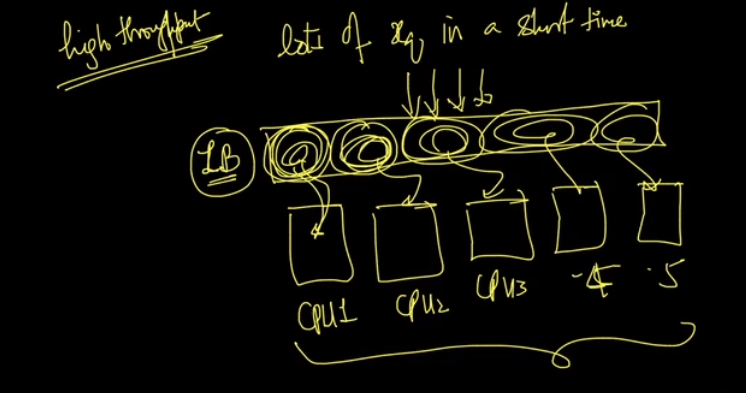
Similarly we can do it for DT/RF’s.

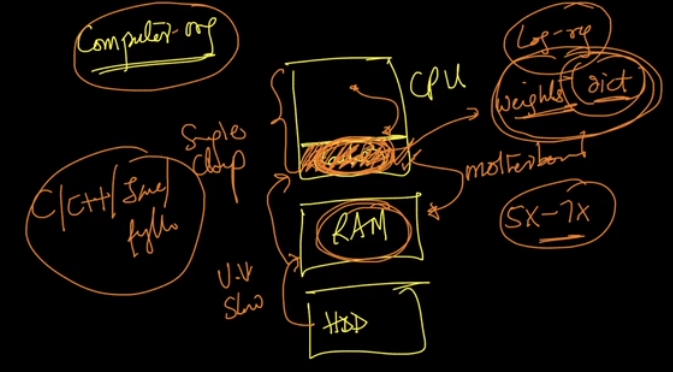


So suppose if we get lots of query in short time. Say companies like google gets thousands of requests in min.

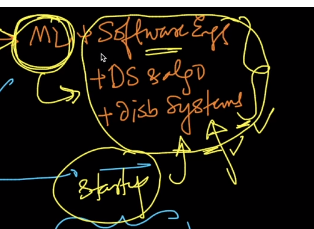
SO to handle there is something called Load Balancer and it just splits the random query points to different CPU’s and if the weights or our If-else statements are stored in cache

Instead of RAM or Hard disks then it can improve performance 5x-7x depending upon the cache size we have.

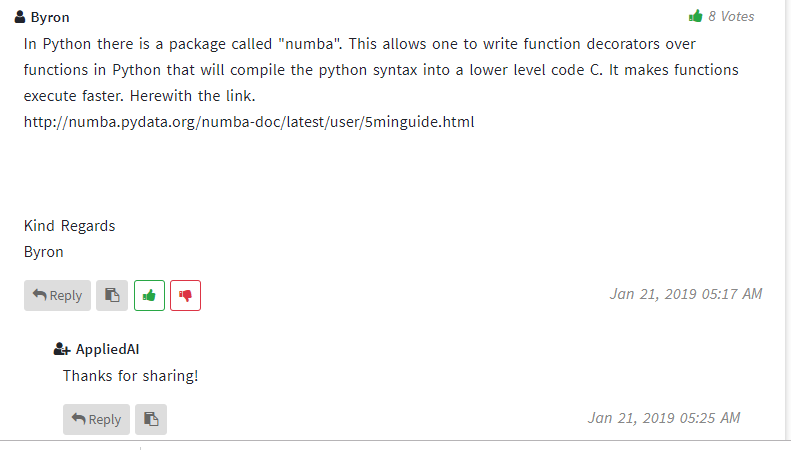




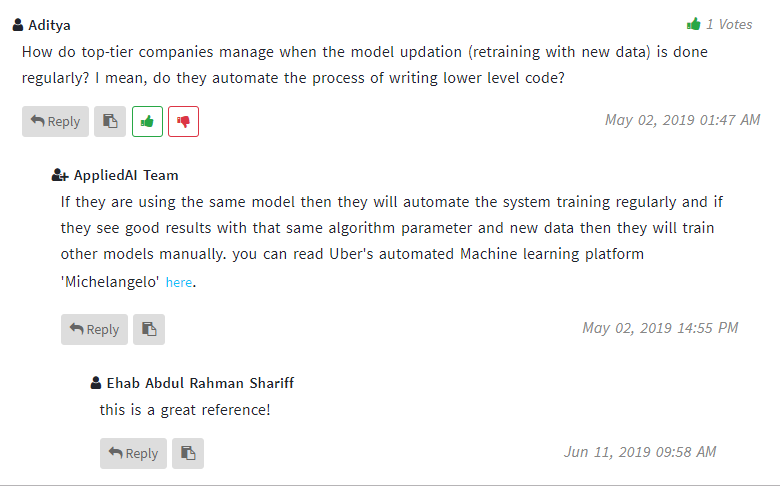
Productionizing model is not only about ML but it involves lot of software engineering, DS, Distributed Systems.



**Comments:**



<http://numba.pydata.org/numba-doc/latest/user/5minguide.html>



<https://eng.uber.com/michelangelo/>