

RAPIDS Taxi Fares: Transcript

One of the data sets that is very familiar to data scientists is the “taxi cab”. It’s a Kaggle data set. We actually brought it into OmniSci so we can visualize -- and this is kind of the business user. We’ll go into a data science workflow but this is indicating how you can see and visualize data in your business.

We have the New York City taxi data and this is points on the map that are indicating drop off locations. The **size** of the point on the map is indicating how long the trip is, and the **color** of the dot is the tip amount from zero all the way up to the max value with zero being blue and the max value being yellow.

You can see that there is a lot of blue [indicating zero tip]. Once you explore you realize that the only way that tips are being tracked is through credit card transactions. So the average tip amount that is displayed here on the dashboard is not a true indicator of what tips are. You can quickly just click on “credit” and it will just show credit card transactions.

So this is a GPU accelerated analytics program based upon a GPU accelerated database which is MapD technology.

You can further analyze the data, if maybe you wanted to know where you needed to have larger vehicles for 9 passengers, for example, or more than 6 passengers. So you can just click on 9 passengers and you can see where those trips were for larger groups. You can include 8 passengers because that still requires a larger vehicle and it brings more data onto the map and into the dashboard. It’s very fast and a great way to allow data exploration across the enterprise and you don’t have to be a data scientist.

The next part of the workflow is what if you wanted to predict taxi fares. Maybe when somebody submitted a taxi fare request on an application (like an app on their phone) you wanted to give them an estimate. And you wanted the estimates to be based upon real taxi fares that happened at the beginning of the month. So that’s where the data scientist comes in and they can use RAPIDS to do those predictions.

OmniSci has Jupyter integration into the application now so you can open up the Jupyter notebook right within their application -- we already have it open here. You can run RAPIDS workflows inside of here. What’s great is you don’t have to take it from that GPU accelerated database that I mentioned back into CSVs. You can integrate with their python API and it can be pulled into GPU memory from their database and everything, including ETL, as well as prediction, can happen all on the same GPU memory.

This is the workflow -- typical RAPIDS, where you go in and you import in your libraries, you connect to the data source, and it brings in the data, shows you the data. The next thing you do is you inspect and you clean up the data. For example, maybe you have columns from 2014 -- they were named a certain way, but they did not map the same way in 2015. Well, through RAPIDS you can do that - you can concatenate fields, you can do data cleanup. And that's what we're doing within the python API, here.

This is the cleaning up data phase. The next thing, after you run your data clean up, you can add new columns or new interesting features to that data. For example, if I didn't really have a way to understand if the taxi fare was on a weekend, or if I wanted to know what day of the week it was without having to read it from a date field, for example. We can create that data, and in this case we did and we added it to the end of the CSV file. We also used a distance calculator so we could see the distance of the trip itself.

The next workflow is to pick a training set. So you go in and pick your training set and it's going to use basically 75% of the data to do training on and then it's going to use 25% of the data to test your model. So we go ahead and create that training set and we do the training. It uses XGBoost to do the training. And lastly, we test our model against that 25%. We calculate a root mean error which is basically a standard deviation to see how close we can get [between] the predictions [and] the ground truth. The end result was the taxi fare was within a \$2 estimate which I think is very fair if you were going to provide that to taxi riders.