The Need for Speed: How the Auto Industry Accelerates Machine Learning with Visual Analytics

Zach Izham, VW | Aaron Williams, MapD
March 27, 2018
Introductions

Asghar Ghorbani  
Data Scientist

Zach Izham  
Legend

Aaron Williams  
VP of Global Community

@ghorbani_asghar  
asghar.ghorbani@volkswagen.de  
/in/aghorbani/  
/a-ghorbani

@drizham  
zach.izham@volkswagen.de  
/in/dr-zach-izham-02090b5/  
/drizham

@_arw_  
aaron@mapd.com  
/in/aaronwilliams/  
/williamsaaron

slides: https://speakerdeck.com/mapd/
Agenda

A Real World Problem: Churn
• Partial Dependency Analysis - An Accelerated Review
• A Complete Machine Learning Pipeline
• Demo: Data Engineering + Training + Predictive Analytics + Black Box Interrogation

The GPU Data Frame in Action
• GO.ai and MapD

Q&A
“Every business will become a software business, build applications, use advanced analytics and provide SaaS services.”

- Smart CEO Guy
The Evolution of Competitive Data

Collect It

Make it Predictive

Make It Actionable
Partial Dependency Analysis

An Accelerated Review
Example: Partial Dependency

Assume the following example:

Failure rate of machine component:

Only depends on
hours of work (HoW) of components, $h$

Not (within reason)
Age of components

Assume failure rate, $f(x)$ is only dependent on hours of work and **not** age:

$$f(h) = \frac{1}{1 + \left(\frac{h}{\alpha}\right)^\beta \cdot \text{noise}}$$

Where $\alpha$ and $\beta$ are constants dependent on machine operating conditions

Example Partial Dependency

Simpson’s Paradox

\[ f(h) = \frac{1}{1 + \left( \frac{h}{\alpha} \right)^{-\beta} + \text{noise}} \]
Example Partial Dependency

Impact of Each Variable on Target Value

\[ f_s(X_s) = \mathbb{E}_{X_c} (f(X_s, X_c) | X_s) \]
Generating Data for the Complete State Space

The Failure Rate Generated from the Trained Black Box Model

Investigating the System with the Simulated Data

Partial Dependency Analysis

Impact of each variable on target value  \( f_s(X_s) = E_{X_c} f(X_s, X_c) \)

Collect Data to Build a Model

Generate Data for the Whole State Space

\[ f_s(X_s) = E_{X_c}(f(X_s, X_c) | X_s) \]

\[ f(h) = \frac{1}{1 + \left( h \over \alpha \right)^{-\beta}} + \text{noise} \]

\[ f_s(X_s) = E_{X_c} f(X_s, X_c) \]
Why do We Need GPUs?
Data Size Explosion

Coarse Grid – Small Data
Dense Grid / Data dimensionality – Large Data

Grid resolution 10:

- 1 variable: 10
- 2 variables: $10 \times 10 = 100$
- ...
- 10 variables: $10^{10} = 10,000,000,000$
Analysis Logistics
Data Engineering

Some Background - Objective - Creating the Master Data Frame

Relevant data reside on separate tables and databases
Collecting, cleaning and curating the relevant data

Creating a target variable:

Which cars will not be returning to the garage / service center for service
VW Data Pipeline

Getting Data to the Environment

1. Loading data to MapD database:
   a. Table extracted from database and exported as csv.
   b. mapdql used to create table and import in data.

2. Exploratory Data Analysis:
   a. MapD dashboard used to perform exploratory data analysis (EDA), gives: ‘spatial awareness’ of the data.
   b. Using MapD allows for on the fly investigation on large datasets.
Demo Data Stats

Rough Feeling of Data

Number of rows ~2.2 million
25 relevant columns
2 categorical columns
23 numerical columns
## Hardware Stack

Details of Hardware Setup On Microsoft Azure Instance:
Instance Name: NC24S_V2 Standard

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>vCPUs (Cores)</td>
<td>24</td>
</tr>
<tr>
<td>Storage</td>
<td>448GB</td>
</tr>
<tr>
<td>Data Disks</td>
<td>32</td>
</tr>
<tr>
<td>GPUs</td>
<td>4 Tesla P100-PCIE-16GB</td>
</tr>
<tr>
<td>OS</td>
<td>CentOS</td>
</tr>
</tbody>
</table>
Demo Time!
Machine Learning Pipeline

Personas in Analytics Lifecycle (Illustrative)
MapD is the analytics platform created for GPUs
Advanced memory management
Three-tier caching to GPU RAM for speed and to SSDs for persistent storage

Hot Data
Speedup = 1500x to 5000x
Over Cold Data

Warm Data
Speedup = 35x to 120x
Over Cold Data

Cold Data

GPU RAM (L1)
24GB to 256GB
1000-6000 GB/sec

CPU RAM (L2)
32GB to 3TB
70-120 GB/sec

SSD or NVRAM STORAGE (L3)
250GB to 20TB
1-2 GB/sec

Data Lake/Data Warehouse/System Of Record
The GPU Open Analytics Initiative (GOAI)

Creating common data frameworks to accelerate data science on GPUs

Model Training / Inference

Data Manipulation/Management

Data Interchange (Zero Copy)

Data Processing (Filtering, Joining and Aggregation)
The Time Is Now

Collect It

Make it Predictive

Make It Actionable
ML Examples

- We’ve published a few notebooks showing how to connect to a MapD database and use an ML algorithm to make predictions

  /gpuopenanalytics/demo-docker

- We will also be publishing the notebook from this VW churn example

  /mapd
Next Steps

- community.mapd.com
  Ask questions and share your experiences

- mapd.com/demos
  Play with our demos

- mapd.com/platform/download-community/
  Get our free Community Edition and start playing
Thanks to the whole team!

- Asghar Ghorbani
- Wamsi Viswanath
- Abraham Duplaa