Large-Scale Machine Learning at Verizon

Ashok N. Srivastava, Ph.D.
Chief Data Scientist, Verizon
A Transition in Roles

• Enormous data volumes
• Massive computing infrastructure
• Significant public benefit that touches daily lives
Organizations started to recognize they are operating with blind spots

1 in 3 business leaders frequently make critical decisions without the information they need

53% don’t have access to the information across their organization needed to do their jobs

Factors supporting major decisions:

- Personal Experience: 79% (To a little extent: 31%, To a great extent: 48%)
- Analytics: 62% (To a little extent: 38%, To a great extent: 24%)
- Collective Experience: 52% (To a little extent: 48%, To a great extent: 24%)
Mobile in the United States

Consumers are Mobile

Smartphone penetration
61% 2013
74% will Game
45% will watch Videos
41% will listen to Music

Social Already Mobile

80% CAGR
70% Smartphone consumers using social

Data Traffic Exploding

40%-45% growth projected per year

2,000 Terabytes per day
1 Terabyte (TB) = ~ 1,000 GB

200 TB metadata/day

- What site?
- What page?
- Last site?
- Next site?
- From where?
- What time?
- What app?
- From who?

Sources: eMarketer, Jumptap & comScore, Vision Critical, Verizon internal data
Revenue Streams due to Analytics

- Predictive Analytics
  - Real-time
  - Based on machine learning
  - SaaS + shared revenue

- Discovery Systems
  - Non query-based
  - Based on machine learning
  - SaaS + shared revenue

- Analytics Reports
  - Descriptive, query based
  - Based on domain knowledge
  - SaaS Revenue

- Liberation of Information
  - 2.5 quintillion bytes of data are created daily
  - 90% of existing data has been created in last 2 years
The Orion Cluster at Verizon

VZ Management Center

Web Services & APIs

Reporting and Advanced Analytics

BI Reporting & Dashboards
Domain-Specific Rules Engine
Anomaly & Pattern Detection
Prediction Algorithms
Recommendation Engine
Insight Discovery

Real-time & Batch Processing

Big Data Infrastructure

Hadoop Ecosystem
Massive Parallel Processing
RDBMS
NoSQL DBMS
Apache Projects

Data Feeds

VZW Network, Clickstream, Time, Location Data
PIP and other Enterprise Data
FiOS
Other 3rd party data
Publicly available data

Vertical Solutions

Advertising
Managed Network Services
Cybersecurity (unified network)
Network Health Management
M2M...

Advertising
Managed Network Services
Cybersecurity (unified network)
Network Health Management
M2M...

Vertical Solutions

Advertising
Managed Network Services
Cybersecurity (unified network)
Network Health Management
M2M...

Vertical Solutions

Advertising
Managed Network Services
Cybersecurity (unified network)
Network Health Management
M2M...

Vertical Solutions

Advertising
Managed Network Services
Cybersecurity (unified network)
Network Health Management
M2M...

Vertical Solutions

Advertising
Managed Network Services
Cybersecurity (unified network)
Network Health Management
M2M...
Orion today...
Precision Market Insights
Connecting marketers with consumers, improving engagement and driving audience response
New Challenges for Marketers

The rise of Big Data & Mobile are complicating marketing activities

- Onslaught of 1st and 3rd party information
- Access to key information from channels
- Siloed data stores across internal groups
- Burgeoning channel for customer interactions
- The amount of data from mobile is exploding
- Lack of standards for tracking and targeting

Mobile can be the solution

- Bridge Digital & Physical
- Direct customer relationship
- Context for cross-channel interactions
- Increased efficiency of customer communications

Data

Mobile

• Access to key information from channels
• The amount of data from mobile is exploding
• Lack of standards for tracking and targeting

Mobile can be the solution

- Bridge Digital & Physical
- Direct customer relationship
- Context for cross-channel interactions
- Increased efficiency of customer communications
Mobile + Big Data Solutions for Marketing

App Usage

Clickstream

Location

Demographics

Married 35-40 year old female in DC metro interested in tennis and luxury brands; recently browsed Tory Burch shoes at m.Nordstrom.com and visits their store 3 times per month.

Precision enables better 1:1 understanding of customers across physical and digital contexts to drive more relevant and personalized interactions.
Case Study: Relevant Mobile Ads Drive A Major Sports League

OBJECTIVE
Engage a targeted audience to click on a mobile banner ad, inviting them to purchase and download the NFL Premium App.

PRECISION MEETS THE CHALLENGE
Precision can reach a broad audience across multiple properties and devices, accurately targeting specific mobile user segments.

AUDIENCE
- Known interest in sports
- Previous basic app download
- Have basic app but not premium app

RESULT
Had 64% higher CTR compared to other mobile media properties

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTR</td>
<td>0.38%</td>
<td>0.23%</td>
</tr>
</tbody>
</table>
Adaptive Architecture for Optimal Advertising

Data Driven Customer Model

Automated Decision Maker

Adjustment Mechanism

Observed Consumer Behavior
Automatic Profile Discovery

Given a data matrix $A$, compute a factorization $A = XY$ such that:

$$\minimize_{X,Y} \| A - XY \|_F^2 + \alpha \| X \|_F^2 + \beta \| Y \|_F^2$$

subject to $X \geq 0$, $Y \geq 0$

This is a nonconvex problem but can be solved at scale using an alternating minimization approach. See S. P. Boyd et. al., 2014, and D. Das and S. Das, 2014.
Large-scale Machine Learning with Applications to Connected Machines
Connected Ecosystem Market Size

By 2017, in the U.S., there will be...

- 207 Million Smartphones
- Over 426 Million Ways to Interact with Customers
- 58.3 Million M2M devices
- 161 Million Tablets

Source: eMarketer, August 2013 and Frost & Sullivan 2013

- Healthcare
- Finance
- Energy
- Manufacturing
- Transportation
- Distribution
Telematics / Fleet Management Today

Making Safety and Efficiency Real
Connected Machine Landscape (I)

Car

- Location analytics
- Comparative analytics across vehicle makes/models
- Prognostic and diagnostic vehicle health management (self-healing autonomous systems)

250M cars on the road in the US, with about 20M new cars sold each year.
Connected Machine Landscape (II)

Car + Consumer

- Driver behavior (safety, remote management)
  - Introduces uncertainty
  - Need for personalization
  - Autonomous control
- Consumer behavior (infotainment, advertising, etc.)
  - Information retrieval
  - Relevance

Over 100M mobile subscribers currently on Verizon’s network
Connected Machine Landscape (III)

Car + Consumer + Environment

- Environmental inputs (Traffic, weather, emergency services, etc.)
- Data correlation
- Distributive decision making and optimization

100s of TB of environmental data generated daily
Verizon has core competency in advanced analytics areas including:

- Anomaly Detection
- Diagnostics
- Predictive Analytics
- Time Series Analysis and Forecasting
- Correlations and Association Analysis
- Optimization
Connected Machine Health Management

1. Anomaly Detection
   “Is something different?”

2. Diagnosis
   “What is happening?”

3. Prognosis
   “What will happen next and when?”

4. Mitigation
   “What actions to take?”

- Real- and near-real-time monitoring and anomaly detection
- Comprehensive and real-time view of big data relevant to diagnosis
- Domain rules and expert opinion
- Predictive Analytics
- Domain rules and expert opinion
- Manual intervention or autonomous and self-healing (“intelligent”) systems
- Data and domain rules
Multiple Kernel Anomaly Detection (MKAD)

Primary Source: Switches, routers and other machines

Primary Source: Maintenance Reports

Primary Source: Network connectivity

Continuous

Discrete

Textual

Networks
Optimization problem

One class SVMs training algorithms require solving the quadratic problem

**Dual form**

\[
Q = \min \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j \left( \sum_{\lambda} \beta_\lambda K_{i,j}^\lambda \right)
\]

Subject to:

1. \( \sum_i \alpha_i = 1 \)
2. \( \nu \in [0,1] \), \( \nu \leq \frac{1}{l \nu} \), \( \forall i \)

\( \alpha \): Lagrange multipliers of the primal QP problem

Linear equality constraint

Control parameter

Bounds on design variables
Anomaly scores

Decision boundary is determined only by margin and non-margin support vectors obtained by solving the QP problem

\[ h(\alpha, \beta, f_z, \rho) = \sum_i \alpha_i \left( \sum_\lambda \beta_\lambda K_{i,z}^\lambda \right) - \rho \]

Data points with \( \alpha_k > 0 \) will be the support vectors

**Indicator**

Sign of \( h \): if negative - outlier
if positive - normal

Value of \( h \): degree of anomalousness
The impact of Big Data Innovations...

The addition of Numeric and Text Data in the new MKAD algorithm significantly improves (by as much as 7000 points) the ranking of the monitored anomalies. Execution time increased by approximately 5%.

The Impact of Big Data Innovations...
Anomaly Detection using NASA MKAD Algorithm

Reported Exceedance
Level 3: Speed low at touchdown
Level 2: Flaps at questionable setting at landing
Big Data and Telematics
Telematics

What Can We Build?

**FLEET**
Monitor fleet location, condition and driver behavior

**WORKFORCE**
Align resource capacity with work and direct the right resource to the right location at the right time

**CLOUD ANALYTICS NETWORK**

**WORKFLOW**
Manage work and share updates both internally and externally

**INVENTORY**
Monitor the location and condition of cargo
Connected Car Solutions

After market data collectors (for all modern makes and models)

Built-in data collectors for Mercedes and other brands
1 BILLION MILES

Dallas, TX

United States

Detroit, MI

New York, NY

Chicago, IL

Houston, TX

58% City Driving, 42% Highway
Real-world MPG calculations are sampled from thousands of vehicles, not just a handful.
Driver behavior is analyzed at multiple levels.

Driver behavior can be compared and contrasted to drive from other OEM vehicles.
Driver behavior is analyzed at multiple levels.

Older model years are driven at significantly lower speeds:

More homogeneous driving across model years
Prediction Interval Estimation using Bootstrap Regression

- Ensemble Learning Approach

We have proven that the empirical quantiles of the bootstrap prediction models can be used to consistently estimate the prediction intervals.

S. Kumar and A. N. Srivastava, under submission to NIPS 2014.

1: **procedure** PIEBR($R(n)$, $\alpha$, $x_0$)
2:   Map Step
3:     Build $T$ bootstrap samples $B_t$ from $R(n)$
4:   for each bootstrap sample $B_t$ do
5:     Build regression models $\tilde{y}_{t,r}(.)$
6:     Evaluate models $\tilde{y}_{t,r}(.)$ at $x_0$, $R(n)$
7:   end for
8: Reduce Step
9: Evaluate the bagged estimator $\bar{y}_r(.)$ at $R(n)$
10: Set $Y_r = \{\tilde{y}_{t,r}(x_0)\}$
11: Initialize error sample set $\hat{E}_r = \phi$
12: for each training sample $(x_i, y_i)$ do
13:     Compute error $\hat{e}_{i,r} = y(x_i) - \bar{y}_r(x_i)$
14:     $\hat{E}_r \rightarrow \hat{E}_r \cup \{\hat{e}_{i,r}\}$
15: end for
16: Form set $E_r$ by sampling $T$ times from $\hat{E}_r$
17: Build the set $C_r = E_r + Y_r$
18: Obtain $C_{\alpha,r}$, $C_{1-\alpha/2,r}$:
19:       the $\alpha/2$, $1-\alpha/2$ quantiles of the set $C_r$
20: Return $I_{\alpha,r}(x_0) = (C_{\alpha,r}, C_{1-\alpha/2,r})$
21: **end procedure**
Anomaly Detection Application

S. Kumar and A. N. Srivastava, under review at NIPS 2014.
Acknowledgements + Notes

- The Entire Verizon Big Data and Analytics Team
- My former team at NASA
- Collaborators at Stanford
- We are hiring for junior and senior roles:
  - Data Scientists
  - Machine Learning experts
  - Platform Engineers
  - Analytics and visualization
  - Application Developers
  - Product Managers