Machine Learning and Data Analytics in Finance
With MATLAB

Account Manager
Matthew Harman, Opti-Num Solutions (matthew@optinum.co.za)

Application Engineers
Kirsten Smith, Opti-Num Solutions (Manager)
Talita Viljoen, Opti-Num Solutions
Danielle Winter, Opti-Num Solutions
Nicole Beevers, MathWorks
Agenda

Introduction

- **Working With Data: Building a Credit Score Card**
  - Explore, clean, extract insights from data
  - Scale up to big data

- **Machine Learning: Building a Bond Rating Engine**
  - Train, validate and customize a classification model
  - Scale up to big data

- **Sharing, Scaling, and Deploying Applications for Production**
MATLAB As A Platform: Our Modeling Impact

- Financial Engineer
- Quant Group

- Traders
- Management
- Other groups

- Regulators
- Clients
- Partners
"Robo-advisors are a class of financial adviser that provide financial advice or portfolio management online with minimal human intervention...based on mathematical rules or algorithms."
- Wikipedia
MATLAB As A Platform
Customer Examples

How do you want to achieve your goal?

- Current Portfolio Value
- Target amount
- Duration
- End date

> Change investment objective

Expectation

- Portfolio value
  - Good market
  - Normal market
  - Very bad market

One-time deposit

Periodic deposits

See Trevor’s full presentation here.
Financial Modeling Workflow

Small/Big Data
- Access
  - Files
  - Databases
  - Datafeeds

Machine Learning
- Explore and Prototype
  - Data Analysis & Visualization
  - Financial Modeling
  - Application Development

Deploy
- Share
  - Reporting
  - Applications
  - Production

Scale
- Small/Big Data
  - Files
  - Databases
  - Datafeeds

- Machine Learning
  - Data Analysis & Visualization
  - Financial Modeling
  - Application Development

- Deploy
  - Share
    - Reporting
    - Applications
    - Production

- Scale
  - Small/Big Data
    - Files
    - Databases
    - Datafeeds

- Machine Learning
  - Data Analysis & Visualization
  - Financial Modeling
  - Application Development

- Deploy
  - Share
    - Reporting
    - Applications
    - Production

- Scale
  - Small/Big Data
    - Files
    - Databases
    - Datafeeds

- Machine Learning
  - Data Analysis & Visualization
  - Financial Modeling
  - Application Development

- Deploy
  - Share
    - Reporting
    - Applications
    - Production

- Scale
  - Small/Big Data
    - Files
    - Databases
    - Datafeeds

- Machine Learning
  - Data Analysis & Visualization
  - Financial Modeling
  - Application Development

- Deploy
  - Share
    - Reporting
    - Applications
    - Production

- Scale
  - Small/Big Data
    - Files
    - Databases
    - Datafeeds

- Machine Learning
  - Data Analysis & Visualization
  - Financial Modeling
  - Application Development

- Deploy
  - Share
    - Reporting
    - Applications
    - Production

- Scale
  - Small/Big Data
    - Files
    - Databases
    - Datafeeds

- Machine Learning
  - Data Analysis & Visualization
  - Financial Modeling
  - Application Development

- Deploy
  - Share
    - Reporting
    - Applications
    - Production

- Scale
  - Small/Big Data
    - Files
    - Databases
    - Datafeeds

- Machine Learning
  - Data Analysis & Visualization
  - Financial Modeling
  - Application Development

- Deploy
  - Share
    - Reporting
    - Applications
    - Production

- Scale
  - Small/Big Data
    - Files
    - Databases
    - Datafeeds

- Machine Learning
  - Data Analysis & Visualization
  - Financial Modeling
  - Application Development

- Deploy
  - Share
    - Reporting
    - Applications
    - Production

- Scale
  - Small/Big Data
    - Files
    - Databases
    - Datafeeds

- Machine Learning
  - Data Analysis & Visualization
  - Financial Modeling
  - Application Development

- Deploy
  - Share
    - Reporting
    - Applications
    - Production

- Scale
  - Small/Big Data
    - Files
    - Databases
    - Datafeeds

- Machine Learning
  - Data Analysis & Visualization
  - Financial Modeling
  - Application Development

- Deploy
  - Share
    - Reporting
    - Applications
    - Production

- Scale
  - Small/Big Data
    - Files
    - Databases
    - Datafeeds

- Machine Learning
  - Data Analysis & Visualization
  - Financial Modeling
  - Application Development

- Deploy
  - Share
    - Reporting
    - Applications
    - Production

- Scale
  - Small/Big Data
    - Files
    - Databases
    - Datafeeds

- Machine Learning
  - Data Analysis & Visualization
  - Financial Modeling
  - Application Development

- Deploy
  - Share
    - Reporting
    - Applications
    - Production

- Scale
  - Small/Big Data
    - Files
    - Databases
    - Datafeeds
How big is big?
What does “Big Data” even mean?

“Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them.”

Wikipedia
So, what’s the (big) problem?

- **Traditional tools and approaches won’t work**
  - Accessing the data is hard; processing it is even harder
  - Need to learn new tools and new coding styles
  - Have to rewrite algorithms, often at a lower level of abstraction

- **Quality of your results can be impacted**
  - e.g., by being forced to work on a subset of your data
  - Learning new tools and rewriting algorithms can hurt productivity

- **Time required to conduct analysis**
  - Need to leverage parallel computing on desktop and cluster
Agenda

- Introduction

Working With Data: Building a Credit Score Card
- Explore, clean, extract insights from data
- Scale up to big data

Machine Learning: Building a Bond Rating Engine
- Train, validate and customize a classification model
- Scale up to big data

Sharing, Scaling, and Deploying Applications for Production
Demo: Consumer Credit Risk Modeling

- Create Credit Scorecards
- Determine Probability of Default
- Calculate Expected Loss
### Manual Investing

#### Build a Portfolio
**Per Loan:** $25

#### Filter Loans
- **Loan Term:**
  - 36-month
  - 60-month
- **Interest Rate:**
  - All: 17.47%
  - A: 6.80%
  - B: 9.33%
  - C: 13.36%
  - D: 17.47%
- **Keyword:**
  - Exclude Loans already

#### Loans Available

<table>
<thead>
<tr>
<th>Rate</th>
<th>Term</th>
<th>FICO®</th>
<th>Amount</th>
<th>Purpose</th>
<th>% Funded</th>
<th>Amount / Time Left</th>
</tr>
</thead>
<tbody>
<tr>
<td>C 3</td>
<td>60</td>
<td>700-704</td>
<td>$20,000</td>
<td>Credit Card Payoff</td>
<td>83%</td>
<td>$3,350, 29 days</td>
</tr>
<tr>
<td>C 4</td>
<td>36</td>
<td>675-679</td>
<td>$6,000</td>
<td>Loan Refinancing &amp; Consolidation</td>
<td>92%</td>
<td>$625, 29 days</td>
</tr>
<tr>
<td>C 5</td>
<td>36</td>
<td>675-679</td>
<td>$32,400</td>
<td>Credit Card Payoff</td>
<td>96%</td>
<td>$1,025, 26 days</td>
</tr>
</tbody>
</table>

Showing Loans 1 - 15 of 269

< back   Available: Sign In

Add to Order
Big solutions

Wouldn’t it be nice if you could:

- Easily access data however it is stored
- Prototype algorithms quickly using small data sets
- Scale up to big data sets running on large clusters
- Using the same intuitive MATLAB syntax you are used to
Process Big Data  

**tall arrays**

- New data type designed for data that doesn’t fit into memory

- Lots of observations (hence “tall”)

- Looks like a normal MATLAB array
  - Supports numeric types, tables, datetimes, strings, etc…
  - Supports several hundred functions for basic math, stats, indexing, etc.
  - **Statistics and Machine Learning Toolbox** support
    (clustering, classification, etc.)

Optional Toolbox: (Improves performance)
- Parallel Computing Toolbox
Big Data in MATLAB: From Desktop to Production

ACCESS DATA

More data and collections of files than fit in memory

1

Datastores

• Images
• Spreadsheets
• Tabular Text
• Custom Files
• SQL
• Hadoop (HDFS)

2

PROCESS ON THE DESKTOP

Adapt traditional processing tools or learn new tools to work with Big Data

Tall Arrays
• Math
• Statistics
• Visualization
• Machine Learning

GPU Arrays
• Matrix Math
• Image Processing

Deep Learning
• Image Classification

SCALE PROBLEM SIZE

To traditional clusters and Big Data systems like Hadoop

Tall Arrays
• Math, Stats, Machine Learning on Spark

Distributed Arrays
• Matrix Math on Compute Clusters

MDCS for EC2
• Cloud-based Compute Cluster

MapReduce
MATLAB API for Spark
Summary of Solutions

- Rich visualizations to extract insight quickly
- Easily tune/fit logistic regression models
- Automate reporting
- Build prototypes and scale without recoding

Toolboxes Used:
- Risk Management Toolbox
Agenda

- Introduction
- Working With Data: Building a Credit Score Card
  - Explore, clean, extract insights from data
  - Scale up to big data
- Machine Learning: Building a Bond Rating Engine
  - Train, validate and customize a classification model
  - Scale up to big data
- Sharing, Scaling, and Deploying Applications for Production
Break
Agenda

- Introduction

- Working With Data: Building a Credit Score Card
  - Explore, clean, extract insights from data
  - Scale up to big data

- Machine Learning: Building a Bond Rating Engine
  - Train, validate and customize a classification model
  - Scale up to big data

- Sharing, Scaling, and Deploying Applications for Production
What is Machine Learning?

“[Machine Learning] gives computers the ability to learn without being explicitly programmed” – Arthur Samuel, 1959

Example: Algorithmic Trading

Standard Approach

Hand Written Program

If RSI > 70 then “SELL”
If MACD > SIG and RSI <= 70 then “HOLD”
...

Formula or Equation

\[ Y_{\text{Trade}} = \beta_1 X_{\text{RSI}} + \beta_2 X_{\text{MACD}} + \beta_3 X_{\text{TSMom}} + \ldots \]

Machine Learning Approach

Machine Learning

model: Inputs → Outputs

model = \langle \text{Machine Learning Algorithm} \rangle (factors, trade decision)
Different Types of Machine Learning

- **Supervised Learning**
  - **Classification**: Output is a choice between classes
    - E.g. Generating trading signals
  - **Regression**: Output is a real number (temperature, stock prices)
    - E.g. Forecasting

- **Unsupervised Learning**
  - Discover a good internal representation
    - E.g. Data Mining

- **Machine Learning (RMSE = 0.015)**
## Challenges in Machine Learning

**Hard to get started**

<table>
<thead>
<tr>
<th>Steps</th>
<th>Challenge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access, explore and analyze data</td>
<td><strong>Data diversity</strong></td>
</tr>
<tr>
<td></td>
<td>Numeric, Images, Signals, Text – not always tabular</td>
</tr>
</tbody>
</table>
Demo: Calibrating the Rating System

- Overseeing a portfolio of bonds
- Improve rating engine using machine learning

\[
X = \begin{bmatrix}
WC_{TA} & RE_{TA} & EBIT_{TA} & MVE_{BVTD} & S_{TA} & Industry
\end{bmatrix}
\]

\[
Y = \begin{bmatrix}
\text{Ratings}
\end{bmatrix}
\]

- AA
- BBB
- A
- BBB
- BB
- AA
- A
- BB
- BB
- AA
- AAA
- CCC
Going Beyond Serial MATLAB Applications
Bonus Demo: FX Trading With Machine Learning

- Time-stamped data
- Create factors and response

<table>
<thead>
<tr>
<th>Time</th>
<th>Mid</th>
<th>Bid</th>
<th>Ask</th>
<th>Ret_N</th>
<th>MACD</th>
<th>RSI_N</th>
</tr>
</thead>
<tbody>
<tr>
<td>01:00:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>02:00:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>03:00:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>04:00:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>05:00:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06:00:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>07:00:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**X**

<table>
<thead>
<tr>
<th>Ret_60</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ret_60 (61)</td>
</tr>
<tr>
<td>Ret_60 (62)</td>
</tr>
<tr>
<td>Ret_60 (63)</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>Ret_60(end)</td>
</tr>
<tr>
<td>NaN</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>NaN</td>
</tr>
</tbody>
</table>

**Y**

<table>
<thead>
<tr>
<th>signal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>-1</td>
</tr>
<tr>
<td>NaN</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>NaN</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>NaN</td>
</tr>
</tbody>
</table>

Trading Decision

Trade Signal:
- $Y > s$ -> BUY (1)
- $Y < s$ -> SELL (-1)
- $Y = s$ -> HOLD (NaN)

Where $s = \text{median(Asks} - \text{Bids)}$
What is Deep Learning?

“Deep learning is the use of artificial neural networks that contain more than one hidden layer” – Wikipedia
Bonus Demo: Volatility Forecasting with Neural Networks

- Response is 128-day volatility
- Performance measured by RMSE
Bonus Demo: Volatility Forecasting with Neural Networks

- Layer Recurrent Neural Network
  - Well-suited to time-series problems
  - Can include more than 1 hidden layer (deep learning)
Summary of Solutions

- Many built-in Apps that generate code
- Lots of machine learning functionality
- No programming background needed
- Easily speed up applications with parallel computing

Toolboxes Used:
- Statistics & Machine Learning
- Parallel Computing Toolbox
Challenge
Improve asset allocation strategies by creating model portfolios with machine learning techniques

Solution
Use MATLAB to develop classification tree, neural network, and support vector machine models, and use MATLAB Distributed Computing Server to run the models in the cloud

Results
- Portfolio performance goals supported
- Processing times cut from 24 hours to 3
- Multiple types of data easily accessed

“Aberdeen Asset Management Implements Machine Learning–Based Portfolio Allocation Models in the Cloud

“The widespread use of MATLAB in the finance community is a real advantage. Many university students learn MATLAB and can contribute right away when they join our team during internship programs. In addition, the strong MATLAB libraries developed by academic researchers help us explore all the possibilities of this programming language.”

Emilio Llorente-Cano
Aberdeen Asset Management

Interns using MATLAB at Aberdeen Asset Management.
Agenda

- Introduction

- Working With Data: Building a Credit Score Card
  - Explore, clean, extract insights from data
  - Scale up to big data

- Machine Learning: Building a Bond Rating Engine
  - Train, validate and customize a classification model
  - Scale up to big data

Sharing, Scaling, and Deploying Applications for Production
MATLAB As A Platform
Example: Portfolio Management Dashboard

- What are my optimal allocations?
- Allow the models to be used for decision-making
Why is sharing models so difficult?

Deployment Challenges in the Enterprise

- Licensing requirements
- Accessing large data
- Responding to new requirements
- Supporting multiple use cases
- Scalability
Deploying Applications with MATLAB

1. MATLAB Application
2. MATLAB Compiler
3. MATLAB Runtime

Toolboxes
MATLAB Desktop
End-User Machine

Deploying Applications with MATLAB
Enterprise Deployment and Scalability of MATLAB

- C/C++
- Java
- C# and .NET
- Python
- Excel
- COM Automation Server
- RESTful, HTTP, and WSDL web services
Summary of Solutions

- User-friendly, interactive tools for compiling without recoding
- Share model royalty-free
- One model can have many different interfaces
- Client-server solution for deployment in the enterprise

Toolboxes Used:
- MATLAB Compiler
### MATLAB Strengths for Machine Learning and Big Data

<table>
<thead>
<tr>
<th>Steps</th>
<th>Challenge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accessing, exploring and analyzing data</td>
<td>Data size and diversity</td>
</tr>
<tr>
<td>Preprocess data</td>
<td>Lack of domain tools</td>
</tr>
<tr>
<td>Train models</td>
<td>Time consuming</td>
</tr>
<tr>
<td>Assess model performance</td>
<td>Avoid pitfalls</td>
</tr>
<tr>
<td></td>
<td>Over Fitting, Speed-Accuracy-Complexity</td>
</tr>
<tr>
<td>Use model for predictions</td>
<td>Recoding costly and error-prone</td>
</tr>
</tbody>
</table>
Learn More: Big Data

- MATLAB Documentation
  - Strategies for Efficient Use of Memory
  - Resolving "Out of Memory" Errors

- Big Data with MATLAB

- Tall Arrays in Action
Learn More: Machine Learning

Documentation:

mathworks.com/machine-learning
MATLAB COMPUTATIONAL FINANCE CONFERENCE 2017

28 September  | New York City

Register at mathworks.com/finconf2017
Get Training

Accelerate your learning curve:
- Customized curriculum
- Learn best practices
- Practice on real-world examples

Options to fit your needs:
- Self-paced (online)
- Instructor led (online and in-person)
- Customized curriculum (on-site)
Q&A
Financial Modeling Workflow