On the fly retraining of predictive analytical models using Spark Streaming: An equity-price direction prediction case study

Tijl Carpels & Dirk Van den Poel
About Me

- PhD. candidate @ Ghent University
- Department of Marketing, Faculty of Economics and Business Administration

Projects
- Social fraud prediction & staff allocation
- Virdata - IoT Platform
- Automotive industry - CLV prediction

Tijl.Carpels@UGent.be
Solution
Stock Market Prediction

- Plethora of research focusing on stock market prediction
- Technical trading rules
  - Forecasting the direction of the stock price by analysis of volume and price
- AI models
  - Growing popularity
  - Higher forecasting accuracy
Past Research

- Offline batch processing
- Concept drift
- Data granularity
Efficient Market Hypothesis
Our Approach

- **Big Data use-case**
  - Deployable in reality
  - Learn from each retraining
  - ‘Continuous Application’ (see Spark 2.0)
Flow

- Stock Data
- Monitor performance
- Model Manager
- Batch Model Retraining
- Decisions: Buy/Sell/Hold
- Score & Decide
- model
Demo
Our model

- Predict Buy, Sell and Hold positions
- We replay history: April 2015 until April 2016
- Day trading

<table>
<thead>
<tr>
<th>Time</th>
<th>Price</th>
<th>Difference</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.00</td>
<td>14.00</td>
<td>+1</td>
<td>Buy</td>
</tr>
<tr>
<td>14.00</td>
<td>16.00</td>
<td>-1</td>
<td>Sell</td>
</tr>
<tr>
<td>16.00</td>
<td>14.95</td>
<td>+0.05</td>
<td>?</td>
</tr>
<tr>
<td>Closing time</td>
<td>15.00</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>
Thresholds
Feature Engineering

- Financial metrics / Technical trading rules
  - Simple Moving Average
  - Bollinger Bands
  - Commodity Channel Index
  - Weighted Moving Average
  - Stochastic Oscillator
  - Chaikin Accumulation/Distribution
  - Moving Average Convergence Divergence
  - Relative Strength Index
  - Triple Smoothed Exponential Average
  - William %R
  - Money Flow Indicator

‘Interaction’ Effects
Our Approach

Stock Data → Score & Decide → Model Manager → Monitor performance → Batch Model Retraining

Decisions

Buy/Sell/Hold
Cluster 1
1. Request list of active models
2. Request model specifications of active models
3. Wrangle and Score
4. Calculate performance

Cluster 2
1. Batch retrain model with specified parameters

Diagram:
- Score & Decide
  - model
- Monitor performance
- Model Manager
- Batch Model Retraining
Model Manager

- Database of models
  - RFModel, StandardScaler, feature set

- Automation of data science tasks
  - Time frames (training data)
  - Cut-off values (hold position)
  - Real-time decision making

- Track evolution of models
  - Time periods, features, performance
Demo
Research*

● Model Manager
  ○ Threshold for hold signals
  ○ Training period

*Master dissertation Giselle Van Dongen
Training period

• Training period
  ○ 1 day
  ○ 1 week
  ○ 2 weeks
  ○ 3 weeks
  ○ 1 month
  ○ 2 months
  ○ 3 months
  ○ 6 months
Average Daily Return (t=0.5) (AAPL)
Average Daily Return (t=0.2) (AAPL)
Average Daily Return (t=0.75) (AAPL)
Cumulative Return (t=0.75) (AAPL)
Average Daily Return (t=0.5) (XOM)
Average Daily Return (t=0.75) (XOM)
Cumulative Return (t=0.75) (XOM)
Conclusions

● Design
  ○ Simulation
  ○ Flexible
  ○ Extendible

● Spark
  ○ Both ML and Stream processing
  ○ Spark 2.0 and 2.1

● Research
  ○ Training time span
  ○ Threshold B/S/H
Roadmap

- Determine threshold and retraining strategy
- Incorporate transaction costs
- Add portfolio optimization module (risk management)
- Simulate all S&P100 stocks
- Benchmark results
Thank you for your attention

www.bigdata.UGent.be