Ranking Private Companies using Predictive Indicators
Overview

- Background on private equity investing in growth stage companies
- Introduction to applying machine learning to support investment decisions
- Extracting features, modeling and validation
- A Bayesian ranking methodology
The Research Goal

To inform investment decisions for growth-stage companies by using quantitative techniques to learn predictive indicators and patterns in historical data.

These indicators and patterns are then used to infer the likelihood of successes and failures of companies for potential investments.
Two Converging Trends: Data and Analysis

- Increasing Amount of Data
- Increasingly Powerful Analytics

- Capital IQ
- DOWJONES
- Thomson Reuters
- CleanTech
- Crunchbase
- Quid

- Google Ventures: $165M
- Correlation Ventures: $100M

- Data Mining
- Algorithmic Trading
- Fraud Detection
Opportunities in Growth Stage Investments

- Quantitative Analysis
  - Best assisting at the top of the opportunity funnel
The Analysis Approach

**Hypothesis**

**A:** Mine for predictive indicator, then interpret the finding

**B:** Start with a hypothesis, then quantitatively validate or invalidate it

**Techniques**

**A:** Supervised Learning using labeled data set

**B:** Unsupervised Learning using unlabeled data set

**Data**

**A:** Labels of successes or failures

**B:** Quantifiable company characteristics that can be consistently measured
Big Data companies have outperformed their respective markets and have created competitive advantage.

Percent, 10-year CAGR (1999 – 2009)

Revenue

- Grocers: 6% for Big data leader, 12% for Other competitors
- Online retailers: -1% for Big data leader, 24% for Other competitors
- Big box retailers: 5% for Big data leader, 9% for Other competitors
- Casinos: 5% for Big data leader, 11% for Other competitors
- Credit cards: 9% for Big data leader, 14% for Other competitors
- Insurance: 9% for Big data leader, 8% for Other competitors

EBITDA

- Grocers: 3% for Big data leader, 11% for Other competitors
- Online retailers: -15% for Big data leader, 22% for Other competitors
- Big box retailers: 2% for Big data leader, 10% for Other competitors
- Casinos: 1% for Big data leader, 12% for Other competitors
- Credit cards: 1% for Big data leader, 9% for Other competitors
- Insurance: 5% for Big data leader, 14% for Other competitors

SOURCE: Bloomberg and Datastream, annual reports, McKinsey analysis
Challenge 1: A Nascent Industry

- **Challenge:** In Cleantech, only a small number of companies have known exits.

  - 100% 3775 clean tech companies identified
  - 94.0% 3550 clean tech companies with investment history
  - 10.9% IPO + IPO Registration + known acquisition success + Bankrupt + Chapter 11 + known acquisition failure

- **Solution:**
  - A technique to infer relative successes and failures of companies based on various observed features.

![Graph showing company valuation over years](chart.png)
Challenge 2: Lack of Objective Measures of Success

**Challenge:** Limited by poor data quality in public data sets
- < 15% of companies have post money valuation for Cleantech in VentureSource

**Solution:**
- Use available data, detect when enough is available to be statistically significant
- Identify missing data points that are most useful to substantiate patterns in the data set
Modeling in a Private Equity Firm

Investment Thesis

- Value chain
- Sector landscape
- Business Model
- Competitive position
- Valuation
- Industry size?
- Winners & losers?
- Unit economics?

Form Statistical Hypothesis

Collect data

Extract Features

Train Model

Evaluate Result

Expected Result?

Understand and Adjust

Usable Model

Yes

No
Extracting Features

Elementary Features

- Number of Companies in a Sector
- Number of Companies in a Geography
- Company Valuation
- Funding Year
- Funding Amount
- Funding Frequency
- Number of investors in a Round
- Total Number of investors in a Company
- Number of Executives
- Investor Location
- Investor Portfolio Size

Advanced Features

- Company Density in a Geographic Region
- Company Valuation Growth
- Difficulty of Funding Environment in Funding Years
- Investment Amount vs Post-Money-Valuation Ratio
- Investor Performance/Return
- CEO Experience (Prior Executive Roles)
- Sector Growth Over Time
- Investor Prestige Over Time

Investor Prestige (Based on Syndication Pattern)
Highly Volatile Sectors in the Cleantech Industry

Company Valuation Growth within its Sector

- **Problem:** How to objectively assess a company amidst large sector volatility?

- **Solution:**
  - Peer-based analysis on company growth
  - Use peer-average to make company performance “invariant” to sector volatility
Normalized Growth within Sector: by Sharpe Ratios

Investment Valuation Growth Normalized to Annual Sector Average

3S Swiss Sustainable Systems
Applied Solar Technologies (India) Pvt. Ltd.
Astropower
BrightSource Energy Inc.
Changzhou Almaden PV Glasses Co. Ltd.
Chaowei Power Holdings Ltd.
ClimateWell AB
Dalian East New Energy Development Co. Ltd.
Daqo New Energy Corp.
Echogen Power Systems Inc.
Evergreen Solar
GT Solar International
Henan Xindaxin Materials Co. Ltd.
Infinia Corp.
Innergy Power
Intersolar
Jiangsu Xiuqiang Glasswork Co. Ltd.
JinkoSolar Holding Co. Ltd.
JIPelec
Liao Ning Oxiranchem Inc.
Miasole
Nanosolar Inc.
Nidecon Technologies Oy
NovaCentrix
OmniPV
Normalized Growth: Comparison Between Sector

Sector Average Annual Growth Rates against Year

- Solar Energy
- Vertical Market Applications Software: Energy Industry
- Biofuels/Biomass

Annual Growth Rates against Year

Assessing the Investor Prestige

**Problem:** How to objectively assess an Investor’s prestige for investment decisions?

**Solution:**
- **Insight:** Best evaluator of an investment are peer-investors
- Analysis with the Page-rank algorithm used in Google Search index
- An investor gets “prestige points” when another investor makes a follow-on investment
# Investor Prestige is Sector-Specific

<table>
<thead>
<tr>
<th>Rank</th>
<th>Investor</th>
<th>Prestige in Solar Sector</th>
<th>Prestige in Energy and Resources Sector</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Nth Power LLC</td>
<td>1.11</td>
<td>13.63</td>
</tr>
<tr>
<td>2</td>
<td>Draper Fisher Jurvetson</td>
<td><strong>3.93</strong></td>
<td><strong>11.56</strong></td>
</tr>
<tr>
<td>3</td>
<td>Khosla Ventures</td>
<td>2.69</td>
<td>9.82</td>
</tr>
<tr>
<td>4</td>
<td>Kleiner Perkins Caufield &amp; Byers</td>
<td>1.79</td>
<td>8.54</td>
</tr>
<tr>
<td>5</td>
<td>RockPort Capital Partners</td>
<td>2.78</td>
<td>8.09</td>
</tr>
<tr>
<td>6</td>
<td>Chrysalix</td>
<td>1.33</td>
<td>7.28</td>
</tr>
<tr>
<td>7</td>
<td>Zero Stage Capital</td>
<td>8.11</td>
<td>7.20</td>
</tr>
<tr>
<td>8</td>
<td>Innovation Works</td>
<td>4.52</td>
<td>6.49</td>
</tr>
<tr>
<td>9</td>
<td>Arete Ventures</td>
<td>5.62</td>
<td>5.95</td>
</tr>
<tr>
<td>10</td>
<td>Sequoia Capital</td>
<td>0.15</td>
<td>5.89</td>
</tr>
<tr>
<td>11</td>
<td>VantagePoint Capital Partners</td>
<td>1.99</td>
<td>5.31</td>
</tr>
<tr>
<td>12</td>
<td>Polaris Venture Partners</td>
<td>3.80</td>
<td>5.27</td>
</tr>
<tr>
<td>13</td>
<td>New Enterprise Associates</td>
<td>5.10</td>
<td>5.19</td>
</tr>
<tr>
<td>14</td>
<td>NGEN Partners</td>
<td>3.79</td>
<td>4.87</td>
</tr>
<tr>
<td>15</td>
<td>Seed Company Partners</td>
<td>N/A</td>
<td>4.82</td>
</tr>
</tbody>
</table>
Investor Prestige Applied to a Company

Example: Enphase Energy Inc.

<table>
<thead>
<tr>
<th>Round</th>
<th>Investors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Round 1</td>
<td>Applied Venture</td>
</tr>
<tr>
<td>Round 2</td>
<td>Applied Venture, RockPort Capital Partners, Third Point Management</td>
</tr>
<tr>
<td>Round 3</td>
<td>Applied Venture, RockPort Capital Partners, Third Point Management, Madrone Capital Partners, Bay Partners</td>
</tr>
<tr>
<td>Round 4</td>
<td>Applied Venture, RockPort Capital Partners, Third Point Management, Madrone Capital Partners, Bay Partners, KPCB, PCG Asset management LLC</td>
</tr>
<tr>
<td>Round 5</td>
<td>IPO</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Investor</th>
<th>Cleantech</th>
<th>Energy/Electricity Generation</th>
<th>Solar</th>
<th>Photovoltaic</th>
<th>Inverters</th>
<th>Micro-inverters</th>
<th>Involvement (*Lead)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RockPort Capital Partners</td>
<td>8.125</td>
<td>3.966</td>
<td>2.785</td>
<td>1.568</td>
<td>1.279</td>
<td>1.279</td>
<td>2*, 3</td>
</tr>
<tr>
<td>Third Point Management</td>
<td>0.276</td>
<td>0.317</td>
<td>0.321</td>
<td>0.371</td>
<td>1.279</td>
<td>1.279</td>
<td>2, 3</td>
</tr>
<tr>
<td>Applied Ventures</td>
<td>1.93</td>
<td>2.091</td>
<td>2.006</td>
<td>2.807</td>
<td>2.142</td>
<td>2.142</td>
<td>1*, 2, 3</td>
</tr>
<tr>
<td>Madrone Capital Partners</td>
<td>0.961</td>
<td>0.943</td>
<td>0.986</td>
<td>0.876</td>
<td>0.15</td>
<td>0.15</td>
<td>3*</td>
</tr>
<tr>
<td>Bay Partners LLC</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>3</td>
</tr>
</tbody>
</table>
Keyword Taxonomy

- Nodes are keywords:
  - Size represents the number of appearances

- Edges are entries in the intersection matrix:
  - Number represents the number of companies sharing the keyword

- Developed a new algorithm to extract hidden structure:
  - Produced a customized spanning-tree for inferring parent-child relationships to produce a taxonomy
Modeling in a Private Equity Firm

**Investment Thesis**

- Value chain
- Sector landscape
- Business Model
- Competitive position
- Valuation
- Industry size?
- Winners & losers?
- Unit economics?

**Form Statistical Hypothesis**

- Collect data

- Extract Features

- Train Model

- Evaluate Result

**Expected Result?**

- Yes: Usable Model
- No: Understand and Adjust

**Expected Result?**

- Yes: Usable Model
- No: Understand and Adjust
Training Models: Exploring the Feature Space

- **Feature Space:**
  - A space in which we can explore the relationships between data points
  - Green: Success
  - Red: Failure
  - Grey: Unknown

- Feature space created by different feature sets can have discriminating capability for company successes and failures
Supervised Learning Infrastructure

- **Use SVM classifier for training models provided by the e1071 R package**
  - E1071 is an R interface to the libSVM library
  - Provides four types of kernels – linear, polynomial, radial and sigmoid
  - Tuning functionality for optimal epsilon, gamma, cost function etc
  - Scales features to have mean zero and standard deviation of unity

- **Use various performance evaluation approaches**
  - Use 2/3 data for training, 1/3 for testing on out-of-sample data or 70/15/15 for training, validating and testing
  - libsvm provides k-fold validation
  - No direct support for ensemble methods – implemented bagging

- **Use Rattle in R**
  - Provides decision tree (rpart), random forest, ada boosting and k-svm

- **Use the coin package for statistical tests**
# Performance Comparison of Classifiers

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Recall</th>
<th>Specificity</th>
<th>Overall Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Tree (rpart)</td>
<td>37%</td>
<td>34%</td>
<td>29%</td>
</tr>
<tr>
<td>Ada Boost</td>
<td>39%</td>
<td>40%</td>
<td>21%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>45%</td>
<td>33%</td>
<td>22%</td>
</tr>
<tr>
<td>SVM</td>
<td>51%</td>
<td>31%</td>
<td>18%</td>
</tr>
</tbody>
</table>
Performance Comparison of Various Classifiers
Training Model: Analyzing the Feature Space

• Model Training:
  – Develop a perspective, based on available data, what is the characteristics of successes vs failures

• Establishing a Decision Boundary:
  – Develop a hyperplane to separate successes and failures
  – Generalize from a limited set of examples
Bayesian Probabilities for Ranking Companies

\[ P(C = S \mid X > x) = \frac{P(X > x \mid C = S) \cdot P(C = S)}{P(X > x)} \]

- **X** denotes the shortest Euclidean distance of a point in feature space to the separating hyper-plane.
- **C** denotes the binary class - (S)uccess or (F)ailure.
Example: Training the model with two features
Posterior Probabilities: cleantech sector

Total Number of Investors

Total Amount Invested (US MM)

Frequency

Nearest distance to the separating hyperplane

Success

Failure

$P(C = S \mid X > x)$

$P(C = F \mid X \leq x)$
Posterior Probabilities: cleantech

- **Total Amount Invested (US MM)**
  - Number of Financing Rounds

- **Probability**
  - $P(C = S \mid X > x)$
  - $P(C = F \mid X \leq x)$

- **Frequency**
  - Success
  - Failure

- Nearest distance to the separating hyperplane
High Dimensional Feature Spaces

<table>
<thead>
<tr>
<th>Feature</th>
<th>Number of Rounds</th>
<th>Number of Investors</th>
<th>Cleantech Transaction Frequency</th>
<th>Total Money In (US MM)</th>
<th>Last Round Investment (US MM)</th>
<th>Last Round PMV (US MM)</th>
<th>Cleantech Prestige</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Missing</td>
<td>6%</td>
<td>13%</td>
<td>73%</td>
<td>40%</td>
<td>40%</td>
<td>86%</td>
<td>40%</td>
<td>89%</td>
</tr>
</tbody>
</table>
Statistical Significance of Features

- Consider the following feature pair – number of investors and total money in (US$M)
- The red/light blue filled circles indicate the company has been successful/failed
- Estimate the distance of the circles to the separating hyperplane
- How statistically significant is the inferred classification using these set of distances?
Statistical Significance of Features

- $H_0$: observed classes of data (i.e. success or failure) are drawn from the same population of feature pairs

- Rejection of $H_0$ implies that a statistically significant classification of feature pairs exists, i.e. the features pairs can be used to infer classes

- Minimum of 5 data points in each class is required for rejection of $H_0$ at the 1% significance level
Agricultural Sector Analysis: Factor models
Bayesian Probabilities: Agricultural sector analysis
# Risk Factor Analysis: e.g. Absorbent Technologies

## Risk Factors:

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Keyword</th>
<th>xFeature</th>
<th>yFeature</th>
<th>Prediction Accuracy (mean / sdev)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(+) 0.4</td>
<td>agriculture</td>
<td>Total Money In (US$ M)</td>
<td>Total Money In (US$ M)</td>
<td>0.854/0.144</td>
</tr>
<tr>
<td>(.) -0.1</td>
<td>agriculture</td>
<td>Last Round Investment (US$ M)</td>
<td>Total Money In (US$ M)</td>
<td>0.852/0.151</td>
</tr>
<tr>
<td>(.) 0.0</td>
<td>agriculture</td>
<td>Number of Investors</td>
<td>Total Money In (US$ M)</td>
<td>0.848/0.140</td>
</tr>
<tr>
<td>(.) -0.1</td>
<td>agriculture</td>
<td>Total Money In (US$ M)</td>
<td>Last Round Investment (US$ M)</td>
<td>0.846/0.153</td>
</tr>
<tr>
<td>(.) -0.2</td>
<td>agriculture</td>
<td>Last Round Investment (US$ M)</td>
<td>Last Round Investment (US$ M)</td>
<td>0.848/0.148</td>
</tr>
<tr>
<td>(.) -0.1</td>
<td>agriculture</td>
<td>Number of Investors</td>
<td>Last Round Investment (US$ M)</td>
<td>0.850/0.144</td>
</tr>
<tr>
<td>(.) -0.5</td>
<td>agriculture</td>
<td>Number of Investors</td>
<td>Total Rounds</td>
<td>0.845/0.167</td>
</tr>
<tr>
<td>(.) 0.0</td>
<td>agriculture</td>
<td>Total Money In (US$ M)</td>
<td>Number of Investors</td>
<td>0.850/0.144</td>
</tr>
<tr>
<td>(.) -0.1</td>
<td>agriculture</td>
<td>Last Round Investment (US$ M)</td>
<td>Number of Investors</td>
<td>0.849/0.141</td>
</tr>
<tr>
<td>(.) -0.5</td>
<td>agriculture</td>
<td>Total Rounds</td>
<td>Number of Investors</td>
<td>0.854/0.158</td>
</tr>
<tr>
<td>(++) 0.6</td>
<td>cleantech</td>
<td>Last Round PMV (US$ M)</td>
<td>Total Money In (US$ M)</td>
<td>0.849/0.047</td>
</tr>
<tr>
<td>(.) -0.1</td>
<td>cleantech</td>
<td>Last Round Investment (US$ M)</td>
<td>Total Money In (US$ M)</td>
<td>0.802/0.039</td>
</tr>
<tr>
<td>(+) 0.4</td>
<td>cleantech</td>
<td>Last Round PMV (US$ M)</td>
<td>Cleantech Transaction Frequency</td>
<td>0.864/0.057</td>
</tr>
<tr>
<td>(++) 0.6</td>
<td>cleantech</td>
<td>Total Money In (US$ M)</td>
<td>Last Round PMV (US$ M)</td>
<td>0.851/0.047</td>
</tr>
<tr>
<td>(+) 0.4</td>
<td>cleantech</td>
<td>Cleantech Transaction Frequency</td>
<td>Last Round PMV (US$ M)</td>
<td>0.865/0.059</td>
</tr>
<tr>
<td>(++) 0.6</td>
<td>cleantech</td>
<td>Last Round PMV (US$ M)</td>
<td>Last Round PMV (US$ M)</td>
<td>0.851/0.043</td>
</tr>
<tr>
<td>(+) 0.2</td>
<td>cleantech</td>
<td>Last Round Investment (US$ M)</td>
<td>Last Round PMV (US$ M)</td>
<td>0.856/0.045</td>
</tr>
</tbody>
</table>
Bayesian Probabilities with Multiple Models

\[ P(C = S | X > x, Y > y) = \frac{P(X > x, Y > y | C = S) \cdot P(C = S)}{P(X > x, Y > y)} \]

\(X\) and \(Y\) denote the shortest Euclidian distances of a point in each of the feature spaces to the embedded separating hyper-planes
Bayesian Probabilities with Multiple Models

Posterior probability of success

Posterior probability of failure
## Ranking of Companies in the Agriculture Sector

**agriculture (97)**

Top level links: [Glossary] [Keywords]

<table>
<thead>
<tr>
<th>Sector analysis: Factor Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>Consumo em Verde Biotecnologia das Plantas</td>
</tr>
<tr>
<td>Jiangxi Tianren Eco Industry</td>
</tr>
<tr>
<td>SweTree Technologies AB</td>
</tr>
<tr>
<td>Guangxi Fenglin Group</td>
</tr>
<tr>
<td>Vitals I/S</td>
</tr>
<tr>
<td>Safer Agro-chem</td>
</tr>
<tr>
<td>GAT Microencapsulation</td>
</tr>
<tr>
<td>Absorbent Technologies</td>
</tr>
<tr>
<td>Agrauxine SA</td>
</tr>
<tr>
<td>Emerald BioAgriculture</td>
</tr>
<tr>
<td>Sn Biotech Laboratories India</td>
</tr>
<tr>
<td>Mycotech</td>
</tr>
<tr>
<td>Idetek</td>
</tr>
<tr>
<td>BotanoCap</td>
</tr>
<tr>
<td>Force A SA</td>
</tr>
</tbody>
</table>
Summary and Future Directions

• Using predictive indicators to rank private companies in a sector is a novel approach for supporting the top-of-funnel diligence process

• The absence of historical company performance data is the primary challenge to enabling this approach

• In the absence of historical valuation data, various features can be selected and extracted for use in a supervised learning model

• Bayesian posterior probabilities can be used to provide a rank of companies within a particular sector

• Future directions include applying the approach to more mature industries
References

- Bhat, H.S. and D. Zaelit, Predicting private company exits using qualitative data, Proceedings of the 15th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD2011), Shenzhen, China, May 2011.


Bagging (slide courtesy of Nikunj Oza)

Test Case
Deviation from the Strategy of Foundation Capital
Deviation from the Strategy of DFJ
Deviation from the Strategy of Khosla Ventures

Sector Classification of Leading Investors' Portfolios: Distance from the Portfolio of Khosla Venture
Relationships Between Investment Styles

Draper Fisher Jurvetson (DFJ) - 44.0

Quercus Trust / David Gelbaum (Private investor) - 18.0
Harris and Harris Group Inc - 15.0
Intel Capital - 20.0
Firelake Capital Management LLC - 12.0
21 Ventures LLC - 11.0
Imperial Champlain Capital Partners - 38.0
Element Partners - 11.0
VantagePoint Ventures - 11.0
PGF Ventures - 9.0
Chevron Technology Ventures (CVV) - 10.0
Applied Ventures - 10.0
Advanced Technology Ventures - 10.0
Kleiner Perkins Caulfield & Byers - 15.0
International Finance Corporation - 12.0
Morgan Stanley Group Inc - 11.0
Braemar Energy Partners LP - 10.0
Carpenter Partners, LLC - 10.0
Good Energies Sigma Capital Group plc - 9.0

Masdar Abu Dhabi Future Energy Company - 10.0
Areté Corporation - 10.0
Credit Agricole Private Equity - 9.0
OPG Ventures Inc - 10.0

Duff Ackerman and Co (DAG Ventures) - 10.0
Westly Group - 9.0
Siemens Venture Capital GmbH (SVC) - 13.0
Schneider Electric Ventures - 9.0

Khosla Ventures - 36.0

SLK ANALYTICS | MARCH 2012
Cleantech Investments up to Year 2000

- Top sector in “Environment”
  - $1.5B in environment
Cleantech Investments up to Year 2003

- **High growth in:**
  - Energy efficiency
  - Energy/electricity generation
  - Energy storage

Diagram showing industry focused products and services, energy/electricity generation, energy efficiency, environment, cleantech, energy storage, etc.
Cleantech Investments up to Year 2005

- High growth in:
  - Water and water treatment processes
  - Fuel Cell investments
Cleantech Investments up to Year 2008

- Explosive growth in:
  - Alternative fuel
    - Natural Gas
    - Biofuel
  - Energy/electricity generation
    - Solar
    - Wind
Cleantech Investments up to Year 2010

- **Steady growth in:**
  - Energy/electricity generation
  - Industry focused product and services
  - Energy efficiency