From data to decisions: Domain knowledge and the machine learning workforce

NBER Economics of AI Conference, U. Toronto

Prasanna (Sonny) Tambe, Wharton School, U. Pennsylvania
Focus is on the division of labor between technology and domain expertise.
Beyond data science: bundling data + domain

Scientist-Marine Mammal Assessments
Ocean Associates, Inc. - San Diego, CA

Apply Now  Save this job

You can send this company applications via Indeed.

Required experience:

- A minimum of 10 years of combined education and experience. Master’s Degree in related field plus two (2) years’ experience or Ph.D. may be substituted for experience.

Additional required qualifications include:

- Advanced expertise using R programming language for data analysis and visualization, including development of R packages and R Shiny applications.
- Expertise conducting Bayesian statistical analyses.
- Expertise customizing mark-recapture type analyses for photo ID data.
- Expertise conducting simulation analysis of managed natural resource systems (e.g., Management Strategy Evaluation).
- Expertise in diverse statistical methods for analyzing ecological data (e.g., generalized linear mixed models; multivariate techniques such as principle component analysis; machine learning techniques such as random forest; spatial statistics).
- Expertise in quantitative methods for studying wildlife population dynamics (e.g., Leslie matrix).
- Experience identifying fish otoliths and cephalopod beaks to species.

Additional preferred qualifications include:

- Knowledge of sea turtle and marine mammal population biology.
- Knowledge of marine mammal diet ecology.
- Knowledge of biological oceanography of the California Current Marine Ecosystem.
- Experience conducting fieldwork for marine mammals.
- Experience communicating ideas to managers and stakeholders.
- Experience with Wilderness First Aid.
ML and job design

• Tests the notion that as we move from data collection → analysis → prediction: there is a growing class of jobs for which technical skills are bundled along with domain knowledge

• Implications for education and for most aspects of labor market activity (job search, etc.)
  • Subtext: Why do so many MBAs want to learn Python?
Using job listings to track skill trends in the ML workforce

- IT investment has historically been difficult to measure, especially machine learning (ML) or artificial intelligence (AI) investments

- **Job listings** provide a way to track granular changes to the technical skills required from workers

- I analyze the Burning Glass Technologies archive of online job listings which includes 1mm+ listings per month between 2010 to 2016
Strengths and limitations of using job listings as a data source

Limitations:

- Challenging to interpret as a vacancy due to sampling
- Within an ad, skills are listed heuristically
- Job listings can be viewed as being aspirational

Strengths:

- Granular skills information (e.g. TensorFlow, Random Forest)
- Within job title variation in skills, i.e. how skills are bundled
- Mapping of skills to job (not skills to worker), and represents where the firm is headed (the gradient)
Interpreting the job listing data:

Job listings can be viewed as a large-scale employer survey asking:

*What skills would you want from a new hire into a specific job title?*

Produces a very large monthly cross-section of responses where employers bear significant costs for providing inaccurate information.
Relevant job skills in the database

• Job skills in database related to **Machine Learning**

  - Machine Learning
  - Decision Trees
  - Predictive Models
  - Deep Learning
  - K-Means
  - Mahout
  - Language Processing
  - Artificial Intelligence
  - Predictive Analytics
  - Data Mining
  - Neural Networks
  - Cluster Analysis
  - Random Forests
  - Support Vector Machines

• **Domain knowledge** mappings taken from the ONET “Work Knowledge” areas (e.g. biology, political science)
Growth in number of listings in database that include ML skills, 2010-2016 (monthly)
Some jobs experiencing rapid growth in ML skills

- Computer Scientists
- Statisticians
- Materials Scientists
- Financial Quantitative Analysts
- Physicists
- Business Intelligence Analysts
- Biological Scientists
- Bioinformatics Scientists
- Social Science Researchers
- Database Architects
- Biostatisticians
- Remote Sensing Scientists
- Electric Installers
- Robotics Engineers
- Economists
Reflects a combination of listings in IT, data science, and functional areas
Among job listings containing ML skills, greater domain knowledge is required from listings coming from functional areas.
This pattern is somewhat unique among information technologies.
Flattens the distribution of technical skills across occupations: data collection.
Flattens the distribution of technical skills across occupations: data analysis
Flattens the distribution of technical skills across occupations: 

**prediction**
Regression framework: Within job-title comparisons

Regressions of the form:

$$ML_i = SKILLS_i + title_i + employer_i + industry_i + \epsilon_i$$

- Allows for measurement of correlations between job skills and workers’ use of new technologies, conditional on job title
- Mitigates some of the concerns about the sampling frame
### Expected correlations with ML and data languages, stats, math, the cloud, and research

**Dependent variable:**

<table>
<thead>
<tr>
<th></th>
<th>Data Lang</th>
<th>Stats</th>
<th>Math</th>
<th>Cloud/Dist</th>
<th>Research</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td>0.148***</td>
<td>0.069***</td>
<td>0.191***</td>
<td>0.185***</td>
<td>0.105***</td>
</tr>
<tr>
<td></td>
<td>(0.019)</td>
<td>(0.003)</td>
<td>(0.014)</td>
<td>(0.011)</td>
<td>(0.014)</td>
</tr>
<tr>
<td>Log(# skills)</td>
<td>0.100***</td>
<td>0.002***</td>
<td>0.074***</td>
<td>0.044***</td>
<td>0.069***</td>
</tr>
<tr>
<td></td>
<td>(0.006)</td>
<td>(0.001)</td>
<td>(0.004)</td>
<td>(0.003)</td>
<td>(0.004)</td>
</tr>
</tbody>
</table>

**Observations:** 12,114 12,114 12,114 12,114 12,114

**R^2:** 0.315 0.060 0.096 0.081 0.082

**Adjusted R^2:** 0.307 0.050 0.087 0.072 0.072

**Notes:** Conditional correlations of ML skills on other skills listed in job ads. 

\[ SKILL_i = ML_i + \log(\text{Num.skills}) + \text{jobtitle}_i + \text{industry}_i + \epsilon_i. \] 
Sample includes listings with ML skills and “matched” listings requiring C++ or Java skills.
Domain knowledge has a stronger relationship with ML than with other data skills

<table>
<thead>
<tr>
<th></th>
<th>SQL (1)</th>
<th>Hadoop (2)</th>
<th>Analytics (3)</th>
<th>SAP (4)</th>
<th>ML (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain</td>
<td>-0.038***</td>
<td>0.011*</td>
<td>0.005</td>
<td>-0.001</td>
<td>0.063***</td>
</tr>
<tr>
<td>(0.011)</td>
<td>(0.006)</td>
<td>(0.004)</td>
<td>(0.003)</td>
<td>(0.004)</td>
<td></td>
</tr>
<tr>
<td>Log(# skills)</td>
<td>0.290***</td>
<td>0.045***</td>
<td>0.019***</td>
<td>0.017***</td>
<td>0.0005</td>
</tr>
<tr>
<td>(0.006)</td>
<td>(0.003)</td>
<td>(0.002)</td>
<td>(0.002)</td>
<td>(0.002)</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>12,026</td>
<td>12,026</td>
<td>12,026</td>
<td>12,026</td>
<td>12,026</td>
</tr>
<tr>
<td>R²</td>
<td>0.257</td>
<td>0.057</td>
<td>0.123</td>
<td>0.059</td>
<td>0.065</td>
</tr>
</tbody>
</table>

Notes: Conditional correlations of ML skills on other skills listed in job ads. 

\[ ML_i = SKILL_i + \log(\text{Num.skills}) + jobtitle_i + industry_i + \epsilon_i \]. Sample includes listings with ML skills and “matched” listings requiring C++ or Java skills.
A correlate: Listings recommend a CS/IS/tech education or domain-relevant college majors (e.g. biology, economics)

Sr Quantitative Finance Analyst
Bank of America 4.04 on 22,329 reviews - New York, NY
10261 (Battery Park area)

Apply On Company Site  Save this job

Required skills
The successful candidate should be a seasoned modeler or validator and meet the following requirements:

- Conducted complete and rigorous independent development and/or validation of models that use machine learning methodologies.
- At least 5 years of work experience at another financial services firm in quantitative research, model development, and/or model validation.
- Graduate degree in mathematics, statistics, computer science, and/or engineering, with a solid knowledge of the banking and finance industry; or possess a graduate degree in finance and/or economics with strong quantitative skills.
- Proficiency in ML platforms/software (e.g., SPM®, Python / sklearn, XGBoost, and R), algorithms, and techniques; and proficient in at least two of the following languages and statistical packages: SAS, SQL, MATLAB, R, VBA, and Python.
- Strong knowledge of financial, mathematical and statistical theories and practices, and a deep understanding of the modeling process, model performance measures, and model risk.
- Strong written and verbal communication skills.

Desired Skills
- Knowledge of risk, underwriting, marketing, valuation, optimization and P&L modeling for consumer banking and lending.
- Coaching experience in a modeling group.
- Ability to manage multiple projects and direct the effort of others.
- Business and operations knowledge and/or experience for auto loans, home loans, credit cards and other products in consumer banking, finance and investments.
These trends are echoed in the larger job listings data

<table>
<thead>
<tr>
<th></th>
<th>CS/IS (1)</th>
<th>Science (2)</th>
<th>Business (3)</th>
<th>Economics (4)</th>
<th>Either (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td>-0.109***</td>
<td>0.022***</td>
<td>0.064***</td>
<td>0.018***</td>
<td>0.104***</td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.004)</td>
<td>(0.009)</td>
<td>(0.003)</td>
<td>(0.016)</td>
</tr>
<tr>
<td>Log(# skills)</td>
<td>-0.001</td>
<td>-0.00001</td>
<td>0.004</td>
<td>-0.004***</td>
<td>0.017**</td>
</tr>
<tr>
<td></td>
<td>(0.005)</td>
<td>(0.002)</td>
<td>(0.004)</td>
<td>(0.001)</td>
<td>(0.007)</td>
</tr>
<tr>
<td>Observations</td>
<td>4,101</td>
<td>4,101</td>
<td>4,101</td>
<td>4,101</td>
<td>4,101</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.432</td>
<td>0.134</td>
<td>0.533</td>
<td>0.047</td>
<td>0.095</td>
</tr>
</tbody>
</table>

**Notes:** Conditional correlations of machine learning skills on degrees listed in job ad. $Degree_i = ML_i + jobtitle_i + industry_i + \epsilon_i$. Sample includes listings with machine learning skills and “matched” workers requiring C++ or Java skills.
Key points

- Suggests that hybrid workers may have a particularly important role to play in the integration of ML tools into business domains

- Flattens the distribution across occupations, continues a trend towards the digitization of new job categories

- Potentially broadens the locus of technical investment within organizations

Comments are very welcome: tambe@wharton.upenn.edu