Roles in DS

Markus Lippus
What we do?

Machine learning and AI development agency

We help companies make sense of their data and turn complex data into insights and profit.
MindTitan’s Services

- Text analysis
- Image recognition
- Recommender engines
- Time-series predictions
- Process optimization
We work with global companies

Industries:
Healthcare, Telecoms, Software, Pharmaceuticals
Data Scientist

Salesman

Software Engineer
Our initial team

- CEO/sales/marketing....
- DS/SE/DE/Analyst/Sales engineer/Sysadmin.....
- DS/SE/DE/Analyst/Sales engineer/Sysadmin.....
Whom we hired

1. Data scientists
2. More data scientists
3. A single DE
Whom we need now

1. Data Engineers
2. Software engineers
3. Sales engineers
Lifecycle of a DS project

- Project inception
- Task specification
- POC development
- Pilot
- Development / Maintenance
<table>
<thead>
<tr>
<th>Decisions</th>
<th>ML task</th>
<th>Value Propositions</th>
<th>Data Sources</th>
<th>Collecting Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>How are predictions used to make decisions that provide the proposed value to the end-user?</td>
<td>Input/output to predict type of problem.</td>
<td>What are we trying to do for the end-user(s) of the predictive system? What objectives are we serving?</td>
<td>Which raw data sources can we use (internal and external)?</td>
<td>How do we get new data to learn from (inputs and outputs)?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Making Predictions</th>
<th>Offline Evaluation</th>
<th>Features</th>
<th>Building Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>When do we make predictions on new inputs? How long do we have to featureize a new input and make a prediction?</td>
<td>Methods and metrics to evaluate the system before deployment.</td>
<td>Input representations extracted from raw data sources.</td>
<td>When do we create/update models with new training data? How long do we have to featureize training inputs and create a model?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Live Evaluation and Monitoring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methods and metrics to evaluate the system after deployment and to quantify value creation.</td>
</tr>
</tbody>
</table>
Making estimates

“How long does it take to implement a paper and achieve 90% accuracy on a vaguely related data set?”
Handling the infrastructure for big data
Just handling large amounts of data

Critical point at 50-100GB

Please learn SQL
Data preparation

- Python is bad at parallelization
- Data scientists are bad at coding

- Learn to write effective code
- Learn to assess computational complexity
- There’s a LOT of communication to do
Algorithm selection and testing

1. Entire Data Set
2. Polynomial Features
   - Pipeline operators modify the features
3. PCA
   - Modified data set flows through the pipeline operators
4. Combine Features
5. Recursive Feature Elimination
6. Random Forest Classifier

Multiple copies of the data set can enter the pipeline for analysis.
Algorithm selection and testing
Algorithm selection and testing
Validation

Experiment!

- Offline
- Online
Validation

- Learn statistics
  - Interpreting results
  - Taking samples
  - Understanding mistakes
Deployment

1. Problem owner
2. Data Scientist
   - Computing cluster
   - Data Science workstations
3. ML models
4. Software Engineer
   - APIs
   - Dashboards
Managing pipelines

1. Control Plane Data
2. Radio Metrics
3. Customer info

Data Scientist -> ML models -> APIs -> Dashboards

Google cloud platform

OR

On-premise hadoop cluster
Building data collection software

CS chats and labels conversations

Data Scientist feedback and guidance for improving labelling

Labelled chats stored for training models

2nd and 3rd opinion from verified specialists
Ongoing work

1. Business leader
   - Understands the business value
   - Communication between parties
   - Setting goals

2. Problem owner
   - Understands the problem
   - Validates results
   - In sync with business

3. Data scientists
   - Prepares the data
   - Mines the data
   - Creates models
   - Evaluates the results
   - Visualizes results

4. Data engineers / Infrastructure
   - Data accessibility
   - Data quality
   - Data usability
Roles in DS

- Data scientist
- Data engineer
- Software Engineer with DS inclinations
- Sales engineer
- Machine learning engineer
- Data science lead

Salesman

Sysadmin

Data scientist

Project manager

Software engineer

MINDTITAN
The end
markus@mindtitan.com