A Day in the Life of a Data Scientist in an AI Company

Francesca Lazzeri & Jaya Mathew

@frlazzeri @mathew_jaya
Agenda

• What is AI and Why is it so important?
• Data, Questions and Metrics
• Understanding the ML Workflow & the Team Data Science Process
• Azure AI
• The Team Workspace
• Model Deployment
• Q&A
What is AI and Why is it so important?
Computer vision and audio processing, for example are able to actively perceive the world around them by acquiring and processing images, sounds and speech. The use of facial recognition at border control kiosks is one practical example of how it can improve productivity.

Natural language processing and inference engines can enable AI systems to analyse and understand the information collected. This technology is used to power the language translation feature of search engine results.

An AI system can take action through technologies such as expert systems and inference engines or undertake actions in the physical world. Auto-pilot features and assisted-braking capabilities in cars are examples of this.

Emerging AI technologies:
- Computer Vision
- Natural Language Processing
- Knowledge Representation
- Machine Learning
- Expert Systems

Illustrative Solutions:
- Virtual Agents
- Identity Analytics
- Cognitive Robotics
- Speech Analytics
- Recommendation Systems
- Data Visualization

Accenture: Why artificial intelligence is the future of growth, April 2016
#TheAIConf @frlazzeri @mathew_jaya
Three major trends are converging

Data
Cloud
Intelligence
From data to decisions and actions

- **Descriptive**
  - Reports

- **Diagnostic**
  - Interactive Dashboards

- **Predictive**
  - Machine Learning

- **Prescriptive**
  - Recommendations & Automation

- What happened?
- Why did it happen?
- What will happen?
- What should I do?

Insight
Digital transformation is driving new business value

Systems of Intelligence

- Engage your customers
- Empower your employees
- Transform your products
- Optimize your operations

@frlazzeri @mathew_jaya
Data, Questions and Metrics
# Being Obsessed with Data

<table>
<thead>
<tr>
<th>Question</th>
<th>Data measures what you care about</th>
<th>Data is accurate</th>
<th>Data is connected</th>
<th>A lot of data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Example:</strong> Predict whether component X will fail in the next Y days</td>
<td><strong>Example:</strong> Identifiers at the level you are predicting, relevant data collected &amp; feature engineering using domain knowledge</td>
<td><strong>Example:</strong> Failures are really failures, human labels on root causes</td>
<td><strong>Example:</strong> Machine information linkable to usage information</td>
<td><strong>Example:</strong> Will be difficult to predict failure accurately with few examples</td>
</tr>
</tbody>
</table>

#TheAIConf @frlazzeri @mathew_jaya
## Asking the right questions

<table>
<thead>
<tr>
<th>Business scenario</th>
<th>Key decision</th>
<th>Data Science question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy forecasting</td>
<td>Should I buy or sell energy contracts?</td>
<td>What will be the long/short-term demand for energy in a region?</td>
</tr>
<tr>
<td>Customer churn</td>
<td>Which customers should I prioritize to reduce churn?</td>
<td>What is probability of churn within X days for each customer?</td>
</tr>
<tr>
<td>Personalized marketing</td>
<td>What product should I offer first?</td>
<td>What is the probability that customer will purchase each product?</td>
</tr>
<tr>
<td>Product feedback</td>
<td>Which service/product needs attention?</td>
<td>What is social media sentiment for each service/product?</td>
</tr>
</tbody>
</table>

#TheAlConf  
[@frlazzeri]  
[@mathew_jaya]
Defining Performance Metrics

1. **Establish a Qualitative Objective**
   - **Example:** Reduce user churn

2. **Translate into a Quantifiable Business Metric**
   - **Example:** Reduce the fraction of users with 4-week inactivity

3. **Quantify the Metric Value Improvement**
   - **Example:** Reduce the fraction of users with 4-week inactivity by 20%

4. **Establish a Baseline**
   - **Example:** Current fraction of users with 4-week inactivity = 60%

5. **Correlation with the Data Science Metric**
   - **Example:** Statistically significant A/B test is a clean way. If this is difficult, compare the values of the metric before and after the solution
Understanding the ML workflow & the Team Data Science Process
Sample ML workflow

Understand Business Goals

Discover /Gather Data
Ingest Data
Understand Data
Transform Data
Create Model

Deploy Model
Monitor/ Maintain Model

Debug, Fix, Enhance, etc.

Respond to changes/lessons

Collaboration & Version Control

Documentation

Build a model

Publish a model

Share Results with Business Owners

Apps
Services
Data Engines

Consume a model
Team Data Science Process
aka.ms/TeamDataScience

Data Sources ➤ Ingest ➤ Prepare ➤ Analyze ➤ Publish ➤ Consume

- Domain expert
- Solution Architect
- Data Engineer
- Data Scientist
- Visualization Expert
- Integration Engineer
- Project Manager
- Executive Sponsorship – IT & Business

#TheAIConf
@frlazzeri
@mathew_jaya
Azure AI
Azure AI

AI apps & agents

- Azure Bot Service
- Azure Cognitive Services

Machine learning

- Azure Databricks
- Azure Machine Learning

Knowledge mining

- Azure Cognitive Search
Sophisticated pretrained models
To simplify solution development

Popular frameworks
To build advanced deep learning solutions

Productive services
To empower data science and development teams

Powerful infrastructure
To accelerate deep learning

Flexible deployment
To deploy and manage models on intelligent cloud and edge
Azure AI

AI apps & agents
- Azure Bot Service
- Azure Cognitive Services

Machine learning
- Azure Databricks
  - Azure Machine Learning

Knowledge mining
- Azure Cognitive Search

@frlazzeri @mathew_jaya
The Team Workspace
The team workspace

Workspace

- compute
- experiments
- data stores
- models
- images
- deployment
from azureml.core import Workspace
ws = Workspace.create(name='Demo',
                      subscription_id='12345678-1234-1234-a0e3-b1a1a3b06324',
                      resource_group='Contoso',
                      location='eastus2')
Model Deployment

**MNIST Dataset**

![21]
Azure Machine Learning deployment workflow
Step 1: Register model

models
Step 1: Register model

```python
# register model
model = run.register_model(model_name='sklearn_mnist', model_path='outputs.sklearn_mnist_model.pkl')
print(model.name, model.id, model.version, sep='\t')
```
Step 2: Register image

**Models**
- Scoring file.py
- Python Environment

**Images**
- Azure Kubernetes Service (AKS)
- Azure Container Instance (ACI)
a. Create scoring script

```python
# %writefile score.py
import json
import numpy as np
import os
import pickle
from sklearn.externals import joblib
from sklearn.linear_model import LogisticRegression

from azureml.core.model import Model

def init():
    global model
    # retrieve the path to the model file using the model name
    model_path = Model.get_model_path('sklearn_mnist')
    model = joblib.load(model_path)

def run(raw_data):
    data = np.array(json.loads(raw_data)['data'])
    # make prediction
    y_hat = model.predict(data)
    return json.dumps(y_hat.tolist())
```
b. Create environment file

```python
from azureml.core.conda_dependencies import CondaDependencies

myenv = CondaDependencies()
myenv.add_conda_package("scikit-learn")

with open("myenv.yml","w") as f:
    f.write(myenv.serialize_to_string())
```
c. Create configuration file

```python
from azureml.core.webservice import AciWebservice

aciconfig = AciWebservice.deploy_configuration(cpu_cores=1,
                                             memory_gb=1,
                                             tags={'data': 'MNIST', 'method': 'sklearn'},
                                             description='Predict MNIST with sklearn')
```
Step 3: Deploy image

- **models**
  - Scoring file.py
  - Python Environment

- **images**
  - Azure Kubernetes Service (AKS)
  - Azure Container Instance (ACI)

- **Deployment**
from azureml.core.webservice import Webservice
from azureml.core.image import ContainerImage

# configure the image
image_config = ContainerImage.image_configuration(execution_script="score.py",
                                                  runtime="python",
                                                  conda_file="myenv.yml")

service = Webservice.deploy_from_model(ws, name='sklearn-mnist-svc',
                                        deployment_config=aciconfig,
                                        models=[model],
                                        image_config=image_config)

service.wait_for_deployment(show_output=True)
Model Deployment Summary

- **models**
  - Scoring file.py
  - Python Environment

- **images**
  - Azure Kubernetes Service (AKS)
  - Azure Container Instance (ACI)

- **deploy**
References

- Azure Machine Learning Services: https://aka.ms/AMLServices
- Data Science Virtual Machine: https://aka.ms/AzureDSVM
- Team Data Science Process: https://aka.ms/TeamDataScience
Thank you!

Learn more

Francesca Lazzeri & Jaya Mathew

@frlazzeri @mathew_jaya