ICT RESEARCH METHODS FOR MACHINE LEARNING

Petra Heck (Fontys), Ralph Niels (HAN), Jeroen Linssen (Saxion)

8 oktober 2021

INTRO

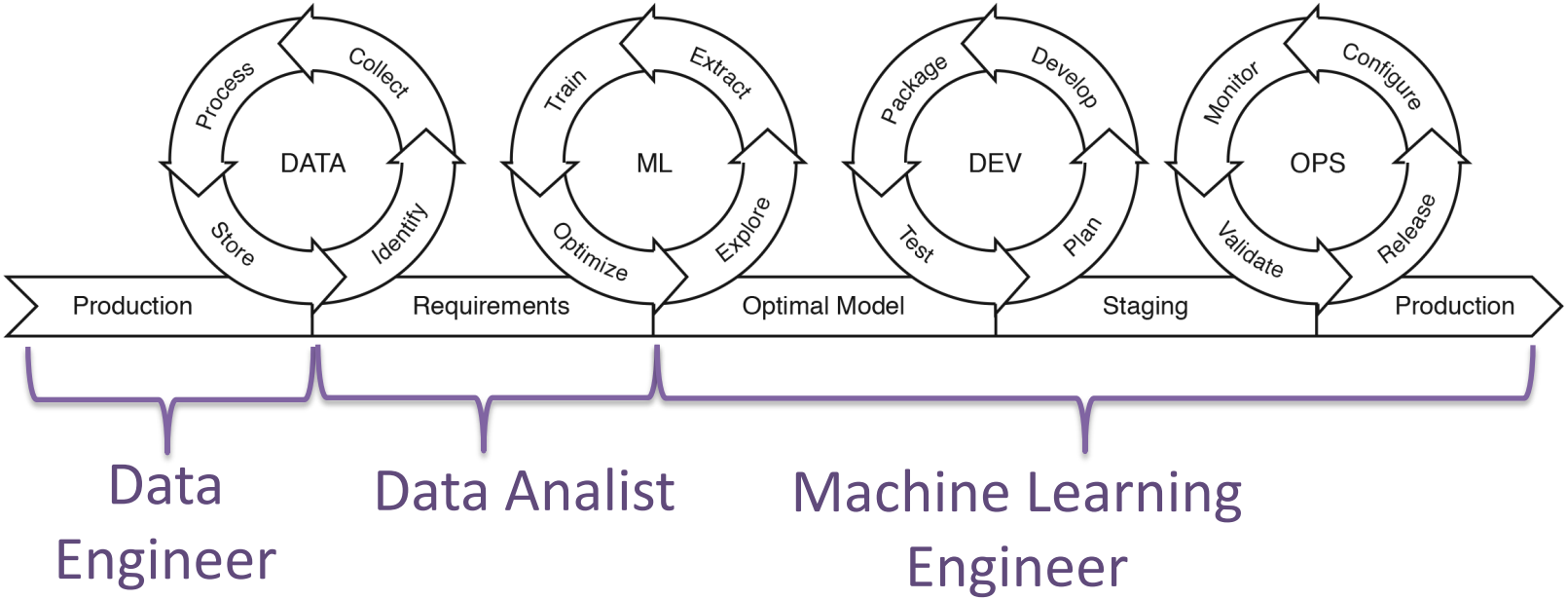
HBO-I addendum Applied Data Science

- ADS = continuum van beroepsrollen en beroepen
- 3 hoofdrollen: data engineer, data analist, ml engineer
- Alternatieve namen voor de rollen inbegrepen
- Per profiel beschreven in HBO-I matrix niveau 3

https://www.hbo-i.nl/wp-content/uploads/2021/03/HBO-i-Domeinbeschrijving-Applied-Data-Science-1.pdf

Fontys

AI engineering en HBO-I ADS rollen



https://fontysblogt.nl/ai-engineering-and-mlops/

ICT & AI Programme @ Fontys – ML Engineer

Semester 1 ICT & Software Engineering Orientation Programming C# (variables, methods and classes)	**Semester 2** ICT & Software Engineering (Web) Application Development Object-Orientation and ASP.NET (SOLID, inheritance)
Semester 3 ICT & Software Engineering Distributed (Web) Applications Full-stack, Java, Angular, REST (Agile, UX, CI/CD)	**Semester 4** ICT & AI Basic Machine Learning Python, Tableau, Excel (Regression, classification, visualisation)

Semester 5 Internship Assignment at company	**Semester 6** ICT & Software Engineering Enterprise Applications Java EE, Docker, Kubernetes (messaging, microservices, non-functionals)
Semester 7 ICT & AI Deep Learning Python, Tensorflow, Keras (NLP, CNN/RNN, Reinforcement learning)	**Semester 8** Graduation Internship Assignment at company

Fontys

https://arxiv.org/abs/2011.01590

ICT Research Methods for Machine Learning

- ML engineering @ HBO gaat over implementeren van AI/ML
  - Zelf maken van AI/ML modellen (data science) is een masterstudie
- ML engineer = Software engineer + Modellen + Data
- Wat is de toolbox van de ML engineer?
  - Welke taken heb je? Welke ontwikkelproces volg je?
  - Welke tools gebruik je? Welke frameworks?
  - ….?
- **Hoe gebruik je de research methods als ML engineer?**

Fontys

CASE: CONTAINER LOGO'S

Opdracht

- Logo herkennen op afzetcontainers (grote vuilcontainers voor een huis)
- Front-end + back-end voor insturen foto's en tonen resultaten (deze afstudeerder maakt alleen backend)
- Eindoplossing moet gebruik maken van Azure Cognitive Services (Vision)

Welke stappen zie je in dit project? Wat zijn de onderzoeksvragen?

Fontys

Onderzoeksvragen

1. Hoe kan Image Classification worden toegepast om containers te kunnen analyseren?
   a) Welke Image recognition API's zijn beschikbaar en hoe kunnen ze worden gebruikt?
   b) Welke kenmerken van een container kunnen herkend worden?
   c) Hoe kunnen kenmerken van een container geanalyseerd worden?
2. Hoe wordt de communicatie tussen de API en de front-end geïmplementeerd/verzorgd?
   a) Welke gegevens moeten er gecommuniceerd worden met de front-end?
   b) Hoe zorgen we ervoor dat de communicatie geauthentiseerd is?

Welke onderzoeksmethodes zou je gebruiken voor deelvraag 1?

Fontys

Research Methods

Exploratory data analysis (ML)

Contents [show]

Why?

Find something interesting in the data, check understanding of the domain or problem space, generate new questions based on the data.

How?

Use descriptive statistics and data visualisation techniques to summarize the data along different axes (features or columns). There is not a pre-defined sequence of actions, instead you determine the path you take through the data (what next to explore?) based on the outcome of the previous step. Stop the exploration once confident that the data has no more secrecies for you.

Ingredients

- A raw dataset (can also be images or documents)
- A domain expert to answers questions about the data
- Creativity to come up with useful data visualisations and exploration paths
- Tools to explore and visualize data (literate programming tools are useful to include explanations in between code)

In practice

Exploratory data analysis (EDA) is a necessary step at the beginning of each data analysis or machine learning project. Next to exploring user requirements you need to explore the data to get yourself familiar with the domain and the problem space.

Model validation (ML)

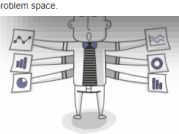
Contents [show]

Why?

Ensure that your model produces results of sufficient quality to base your conclusions on.

How?

While training your model keep in mind how you will ensure that the results obtained from the model will also generalize to cases outside your dataset. Determine the training dataset and the test dataset that you will use. Determine performance measures for your model. Evaluate your models against those measures.



Ingredients

- A (machine learning) model to be validated
- A programming environment to implement the validation measures
- A representative and big enough dataset
- A domain expert to relate your findings to their experience and knowledge

In practice

Standard validation approaches to ensure correctness of models and detect overfitting are cross-validation and bootstrap. Widely adopted correctness measures are accuracy, precision, recall, and AUC. Consider what would constitute a valid model, for example, having an accuracy above a certain threshold.

Data quality check (ML)

Contents [show]

Why?

Ensure that the data you are using is of sufficient quality to base further conclusions on.

How?

Come up with good test cases for your data. Preferably automate those test cases into test scripts. Keep updating the test cases to account for bugs found in the data.

Ingredients

- Understanding of the data (e.g. through Exploratory data analysis)
- A domain expert to answers questions about the data
- A disciplined mindset to cover all important cases
- A critical eye on the validity of your data and your conclusions

In practice

Before you can use a data set in further analyses it is important that you detect incomplete, incorrect, inaccurate, or irrelevant parts of the data. Equivalent to testing code, you also need to test the data and be aware that errors in your conclusions could also stem from errors in the data. Possibly, this may lead to a need for more data, or a conclusion that your research question cannot be answered using the data or that your intended software solution may not meet its requirements.

Model evaluation (ML)

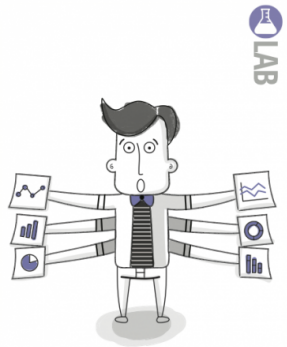
Contents [show]

Why?

Verify the correctness and usefulness of the results of your model with the stakeholders or compare different models with respect to their usefulness.

How?

Translate model performance measures to representations (numbers, pictures, graphs) that are meaningful for the stakeholders. Present these to the stakeholders in a way that makes it easy to collect their feedback. Come up with good "test cases": on which representations do you need feedback and what type of feedback do you need?



Ingredients

- One or more validated (machine learning) models (e.g. through Model validation)
- A tool or application to present model results to the stakeholders
- Understanding of the (business) problem the stakeholders need to address with the model

In practice

If the model results are annotated images, you could build a software application that displays the images and collects end user feedback. If the model results are numbers (e.g. accuracy or probability) you could also use data visualization techniques and tools to present them to the end user. Graphs provide an easy way of comparing different models by showing their results side-by-side.

Fontys

CASE: ANOMALY DETECTION

Opdracht

- Afwijkingen vinden in uitgereikte bio-certificaten aan bedrijven (“fraudedetectie”)
- Front-end + back-end voor tonen afwijkingen, bestaande database met historische data certificaten
- Eindoplossing moet in .NET

Welke stappen zie je in dit project? Wat zijn de onderzoeksvragen?
Hoe passen de ICT Research Methods for ML hierin?

Fontys

Onderzoeksvragen

1. Welke data punten zijn nodig om afwijkingen te herkennen binnen de Skal Biocontrole data?
2. Welke data mag aan de hand van de AVG wet gebruikt worden?
3. Welke anomaly detection tool is het meest geschikt voor dit project?
4. Hoe kan de back-end communiceren met de anomaly detection analyse?
5. Welke back- en front-end wordt er gebruikt tijdens dit project?
6. Hoe wordt een CI/CD pipeline gemaakt in combinatie met anomaly detection tool?
7. Hoe wordt het systeem deployed in combinatie met de anomaly detection tool?

ICT-afstudeerder moet weten hoe met modellen en data te werken

Fontys

AFSLUITING & VRAGEN

Research Methods for ML

Exploratory data analysis (ML)

Contents [show]

Why?

Find something interesting in the data, check understanding of the domain or problem space, generate new questions based on the data.

How?

Use descriptive statistics and data visualisation techniques to summarize the data along different axes (features or columns). There is not a pre-defined sequence of actions, instead you determine the path you take through the data (what next to explore?) based on the outcome of the previous step. Stop the exploration once confident that the data has no more secrecies for you.

Ingredients

- A raw dataset (can also be images or documents)
- A domain expert to answers questions about the data
- Creativity to come up with useful data visualisations and exploration paths
- Tools to explore and visualize data (literate programming tools are useful to include explanations in between code)

In practice

Exploratory data analysis (EDA) is a necessary step at the beginning of each data analysis or machine learning project. Next to exploring user requirements you need to explore the data to get yourself familiar with the domain and the problem space.

Model validation (ML)

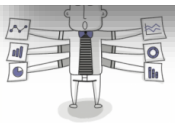
Contents [show]

Why?

Ensure that your model produces results of sufficient quality to base your conclusions on.

How?

While training your model keep in mind how you will ensure that the results obtained from the model will also generalize to cases outside your dataset. Determine the training dataset and the test dataset that you will use. Determine performance measures for your model. Evaluate your models against those measures.



Ingredients

- A (machine learning) model to be validated
- A programming environment to implement the validation measures
- A representative and big enough dataset
- A domain expert to relate your findings to their experience and knowledge

In practice

Standard validation approaches to ensure correctness of models and detect overfitting are cross-validation and bootstrap. Widely adopted correctness measures are accuracy, precision, recall, and AUC. Consider what would constitute a valid model, for example, having an accuracy above a certain threshold.

Data quality check (ML)

Contents [show]

Why?

Ensure that the data you are using is of sufficient quality to base further conclusions on.

How?

Come up with good test cases for your data. Preferably automate those test cases into test scripts. Keep updating the test cases to account for bugs found in the data.

Ingredients

- Understanding of the data (e.g. through Exploratory data analysis)
- A domain expert to answers questions about the data
- A disciplined mindset to cover all important cases
- A critical eye on the validity of your data and your conclusions

In practice

Before you can use a data set in further analyses it is important that you detect incomplete, incorrect, inaccurate, or irrelevant parts of the data. Equivalent to testing code, you also need to test the data and be aware that errors in your conclusions could also stem from errors in the data. Possibly, this may lead to a need for more data, or a conclusion that your research question cannot be answered using the data or that your intended software solution may not meet its requirements.

Model evaluation (ML)

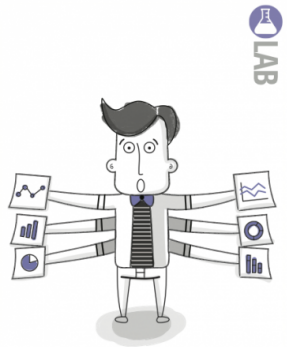
Contents [show]

Why?

Verify the correctness and usefulness of the results of your model with the stakeholders or compare different models with respect to their usefulness.

How?

Translate model performance measures to representations (numbers, pictures, graphs) that are meaningful for the stakeholders. Present these to the stakeholders in a way that makes it easy to collect their feedback. Come up with good "test cases": on which representations do you need feedback and what type of feedback do you need?



Ingredients

- One or more validated (machine learning) models (e.g. through Model validation)
- A tool or application to present model results to the stakeholders
- Understanding of the (business) problem the stakeholders need to address with the model

In practice

If the model results are annotated images, you could build a software application that displays the images and collects end user feedback. If the model results are numbers (e.g. accuracy or probability) you could also use data visualization techniques and tools to present them to the end user. Graphs provide an easy way of comparing different models by showing their results side-by-side.

https://ictresearchmethods.nl/Domain:_Machine_learning