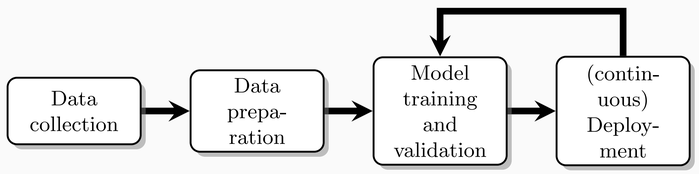
Automation of Machine-Learning Pipelines for Next Best Offer Systems

Maximilian Dierschke, Advisor: Steve Chenoweth
Rose Hulman CSSE department, Ulm University of Applied Sciences

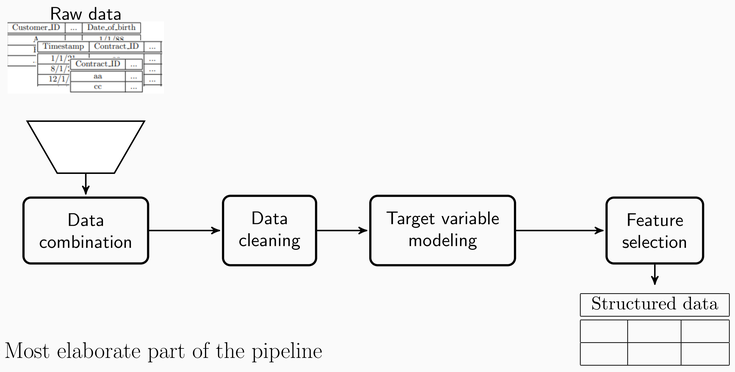
We want to show, that it is possible to automate the creation and deployment of machine learning pipelines, where the user needs to have minimal interaction. This is done for a Next Best Offer system, which is supposed to predict, which product best to suggest to a customer to make a sale.

Machinelearning Pipeline



(Data collection is a problem on its own, and only shown in this graphic for completeness)

Data Preparation



- Most elaborate part of the pipeline
- Structured form for the ML model
- Needs to be reproducible

Model Training

Multiple ML algorithms are well known and are tested in the context of the pipeline.

Deployment

- Versioning for ML model
- Making latest version of ML model available to client

Approach

- Model for insurance data
- Data is distributed over multiple tables
  (main info) in contracts table, kind of policies and other info in other tables
- Multiple policies can be bought each timestep

Data combination

- Combination of 60 tables that needed to be joined with 20 million contracts
- This was possible to be automated with a script, that joined the tables one after the other.
- Optimization not possible because these tables hold information to target variable

Data cleaning

- Depending on the datatype, Null values were replaced with the mode or median
- Duplicates werre removed

Target variable modeling

- Combination of temporal relationship and kind of policies
- In first step consider temporal relationship with self-join (very expensive)
- Then joining the combined vector of policies to main table

ML model

- Predict multiple possible next policies with the data from the most recent contract
- Predicted with a single model (vector of policies)
- **Or** ensamble of binary classificators

Results

- The data combination can be automated:
  - Optimization in this case not possible, since all tables are needed in target variable
- Data cleaning can mostly be automated:
  - Not possible for all datatypes, not all numerical data can be treated as such
- Target variable modeling can't be automated for general probelms
  - Regarding Next Best Offer systems, this might be possible due to similar structure
- Training the ML model can be automated

Model performance

Ensamble:
Subset of 16 policies is used to evaluate the accuracy of the different models. (biary classification for each policy)

	accuracy		
algorithm	average	lowest	highest
decision tree	0.8715	0.7682	0.9989
random forest	0.8594	0.7419	0.9979
neural network	0.7943	0.7216	0.9312

Single model:

- Predicting all policies at the same time in form of a vector
- With different layer sizes for a deep network, this architecture does only approach accuracy of 0.45

Plans before thesis completion

- Testing the created pipeline on other data
- Evaluationg deployment options

Acknowledgements

This thesis is done in collaboration with PROFECT. They provided data, computational resources and great advice.