Maintaining the Data Quality in ML development

Data
Training
Model
Our group at Statistics Finland

• Stats Finland’s departments were reorganized in September 2020

• Digitalization services established, responsible for
  • Managing and assisting with innovations
  • Developing ML-based solutions

• Eight persons, half with ML in job description
ML at Stats Finland

• Limited experience with ML
  • A couple of rudimentary classification models in production
  • A few more cases investigated but have not made it to production

• Finland’s public sector policy: If a cloud service offers the best service benefit and guarantee, it should primarily be selected for new IT-solutions, provided no other barriers exist

• Our aim for 2021
  • Implement a generic platform in Azure allowing classification models to be rapidly developed, put into production and reliably maintained
  • Follow MLOps principles, automate as much as possible
  • Deploy models for two classification cases to production
"Garbage in, garbage out…"

• In the machine learning context, "garbage in..." means that the ML model is only as good as your data

• Therefore, the data, which has been used for training of the ML model, indirectly influence the performance of the whole ML system

• Better understanding of “DS”-process could be the first step in improving the performance of the ML model, how mature is your process?

• After good understanding of DS-process, automate it!
Data

Data Extraction
- Data selected from different sources *

Data Analysis
- Understanding data characteristics

Data Preparation
- Data splitted (training, validation and test data)

Model Training
- Trained model

Model Evaluation
- Metrics to assess the quality of the model

Model Retraining
- Feature engineering usually left outside (not always)

Model Training
- Trained model

Model Validation
- The performance of the model is adequate (or not)

Model Serving
- Model deployed to production

Model Monitoring
- New iteration is invoked if necessary

Feature Engineering
- 2. Understanding data
- 3. Feature selection & creation
- 4. Model training...

* output of the phase
"Quality in, quality out…"

1. **Data Extraction**
   - Select and integrate relevant data

   Key considerations: Ground Truth, Data Relevance, Quantity of Data, Ethics

2. **Data Analysis**
   - Perform Exploratory Data Analysis to understand the available data

   Key considerations: Missing values, Outliers, Unbalanced data, Feature engineering

3. **Data Preparation**
   - Perform Data Cleaning, Data Splitting and Data Transformations

   Key considerations: Categorical encoding, Dealing with skewed data, Scaling, Bias Mitigation
   Feature engineering: Feature extraction, Capturing Feature relationships

(Feature is an attribute used as input for the model to train, or perhaps better definition: individual versioned and documented data column (in a feature store) or even better definition …)
Data Preparation
- Data split (training, validation, and test set)

Test set
- The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset
- Used in model quality evaluation

Training set
- The sample of data used to fit the model

Validation set
- The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration
Circumstances triggering retraining

• ML process easily ends up in struggling with all kind of anomalies: decays, skews, drifts and biases…

• Model performance may be affected by numerous factors

• What then?
  • Identify all those unwanted circumstances and what are the factors behind them
  • Improve the weak parts in your DS-process (improve the degree of maturity of DS-process)
  • Create metrics, monitors and build triggers and *automate*
• Identify necessary roles, and how tasks are related to them

* Concept and picture are both from Gartner
ML workflow could be something like this?

- "DS"-process is in the middle of many other components