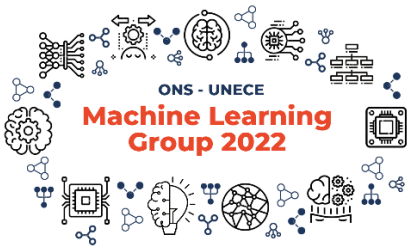

ONS-UNECE Machine Learning Group 2022

Sprint at UK Data Science Campus, Newport, 12-14 July, 2022

Summary Report

Members of the Machine Learning Group 2022 met at the UK Data Science Campus in Newport, UK, 12-14 July for an in-person sprint. The sprint was part of a wider meeting organised by the UK Office for National Statistics entitled: ***Advancing International Collaboration in Data Science and Big Data for Official Statistics***.

The Machine Learning Group 2022 sprint involved three of this year's Theme Groups:

- The web-scraping group aims to develop statistical indicators using web-scraped data and ML methods.
- The model re-training group explores various issues related to model re-training, such as how to decide on model re-training and when to re-train.
- The quality of training data investigates the concept of data quality in the context of machine learning and its effect on the ML model.

All three theme groups have been running regular online meetings since early this year and the International Week sessions aimed to accelerate the work by bringing several members together to meet in person. Twenty-one members attended from 14 different organisations.

On the first day, three theme groups had one joint session where each group presented the main topics under discussion and works done so far. This joint session provided a very useful opportunity to learn more about other theme groups and receive feedback from non-theme group members. The three theme groups then worked in parallel sessions for the following two days.

The participants of the work sessions found in-person discussions very productive, helping expedite the work. The meeting also offered a great opportunity to cross-fertilize among three theme groups as well as receive valuable feedback from outside the theme groups.

The following sections provide a summary of the outcomes of each of the theme group sprints:

Model retraining

Coordinator: InKyung Choi, UNECE.

The theme group's work this year has been focussing on developing a report that could provide a guide for model monitoring and re-training for statistical organisations During the sprint it had four sessions structured around the following topics:

- drifts concepts and metrics to monitor,
- methods to detect drifts,
- organisational aspects to consider,
- ways to retrain the model.

The sessions helped greatly in advancing the discussion around the key terminology and establish common understanding on how different types of drifts are related (for more details about the discussion in each session. Key lessons learned included:

  - The model monitoring and re-training is crucial to use ML in a long term, it seems many practitioners are struggling on their own.
  - There is also no clear role for each actor Intermediary roles between business and data science (functional analyst) and between IT and data science (ML engineer) is important
  - It is still not clear how to measure "drifts" for imagery data and textual data – this discussion will continue after sprint via online meeting

Members found it was very much productive to have in-person discussion compared to online discussions.

As next steps, the group will write up the discussion report and continue discussion and produce report by the end of year. Full notes of the sprint activity can be found on the group's webpage here (UNECE ML Group members only)

Web scraping data

Coordinator: Michael Reusens, Statistics Flanders

In general, the physical sprint was found extremely motivating and valuable by all attending members of the web-scraping activity.

Seven sessions of 90 minutes each were held, comprising:

- 2 sessions with whole ML group: present ongoing work
  - Received valuable feedback on project
  - Learn from other activities
  - Connect with whole ML group
- 2 knowledge sharing sessions within ML group: present individual tools & techniques
  - Learning on methods, big data infrastructure, data science organisation
- 3 sessions: mini-hackaton on applying 2 deep-learning techniques: top2vec/lbl2vec

In terms of the outcomes the group achieved:

  - Very nice early results. Inspiring + promising to develop further.

    - Interesting to see the same methods applied for different purposes
    - Shows the challenges to have good infrastructure to do quick experiments on (install packages, remote connection, enough memory etc.)

The group found it valuable to work physically side-by-side with colleagues (compared to working alone).

In terms of next steps the group will:

- Continue the development of web-scraping based business indicator
- Leverage strengthened connections to continue the collaborative effort,.

More information about the group's work can be found on their web page on the members website here

Quality of Training Data

Coordinator: Marco Puts, CBS Netherlands

During the Sprint, we worked on the representativity of training data. The main question was what representativity means in the case of machine learning. Within the superpopulation theory, we differentiate between the infinite superpopulation (what is the stochastical mechanism that generates the finite population), the population (a realization out of the superpopulation) and a sample (a sample out of the population). Whereas the infinite population is about probabilities, the finite population is about individual cases. The infinite population is based on the concept of exchangeability, whereas the finite population has non-exchangeable items.

When looking at most classification algorithms, it seems that they are much about describing the infinite population than the finite population. We assume that training items are exchangeable (it does not matter which training set we choose) and, for instance, a logistic regression gives the probability of a certain class, which means that it is describing the infinite population instead of the finite population.

Questions we can answer with respect to the finite population are different from questions we can answer with respect to the finite population. The type of questions we answer with regard to the superpopulation is: "what is the probability that …?", whereas questions with regard to the finite population are more related to: "what is the exact number of … in the population?". It may be obvious that official statistics is more about the second question then about the first one. Of course, we often try to answer the question about the amount by using the probability based on the superpopulation and multiplying this with the finite population size. This will in some cases not give the right answer. The finite population is only one realization of the infinite population and therefore introduces a bias. This also holds for the sample with respect to the finite population: the sample has a sampling error, resulting in an error in the model, which gives a bias to the final results based on the machine learning model. Representativeness with respect to machine learning is much about keeping these errors under control.

Experiment

As an experiment, we used the SUSY dataset and tried to predict the number of positive items in the dataset based on 1000 items. Since we know the target variable for the complete population, we could monitor the bias introduced by the different methods. First we used a Horvitz Thomson estimator. The bias based on this estimator was minimal. However, when using a logistic regression model based on the 1000 items and calculating the prediction for the complete population, the estimates of the number of positives was completely off.

Next steps:

Further research needs to be done. We plan to use a jackknife or bootstrap approach on the training set to see if we can calculate the bias introduced by using a small sample from the population. Furthermore, we will deliver a paper describing the difference between the questions answered by a machine learning model and the traditional survey sampling approach and will argue why a survey sampling methodology will bring benefit to the use of machine learning algorithms.

More information about this group can be found on their website on the members website here.