ONS-UNECE Machine Learning Group 2022

UNECE Statistical Division
Workshop on the Modernisation of Official Statistics

Project Report

Alison Baily
International Programme Manager, ONS
23 November 2022

Contact: ML2022@ons.gov.uk
Overview

1. Background
2. Programme Development & Objectives
3. Activity
4. Impact and Lessons Learned
5. Future plans
Background

Machine Learning: “a field of study that gives computers the ability to learn without explicitly being programmed”
Areas with manual, repetitive tasks can be automated with the help of machine learning.

Machine learning – application areas
Objectives

- Platform to facilitate
  - Research to modernise official statistics
  - Building capacity in machine learning
  - Sharing knowledge (data, methods, use cases)

- Community driven
  - Members design programme and provide content
  - Interaction and collaboration is key
  - Every contribution is welcome!

- Public good
  - Open to all official statistical organisations
  - Accessible to different levels of expertise
  - Resources shared with wider community
What we do

Knowledge Sharing
- Monthly meetings with expert presentations
- External engagement at international conferences
- Regular updates of ML news and opportunities

Research Collaboration
- Research projects explore issues from design to implementation
- Findings shared on public website

Capacity Building
- Coffee and Coding sessions
- Learning and training resources
Membership

Public
• Public Website
• Final report + webinar
• Coffee and Coding Sessions

All members
• Monthly meeting
• Newsletter
• Members website
• Catalogue
• Contribute input where possible

Themes
• Research projects
• Study groups
• External presentations
• Regular collaboration
Programme Development and Goals
Goals for 2022

• Focus on moving from proof of concept to production
• Other key areas: C&C, ethics, quality of training data.

• Research collaboration & capability
• A hub for ML news and networking
• Increase membership and active participants
Priority themes for 2022

ML Group 2022 Theme Group Outputs

- Web-scraped Data
- Text Classification
- Imagery Analysis
- AIS Data
- Quality of Training Data
- Model Retraining
- Infrastructure
2022 Timeline

January
Launch workshop
Dubai

March
Theme Group
Activities begin

July
Face to face Sprint,
Newport

November - webinar

December
Final report & close
Activities and Results
Research Collaboration & Knowledge Exchange

- **Web Scraping Data Theme Group**
  - Implementation of experimental statistics using web scraped data for identifying companies active in particular sectors
  - Platform for sharing use cases + discussion best practice

- **AIS Modelling Theme Group**
  - Exploring methods to identify berth areas using ML and AIS data
  - Testing methods at a larger scale + collating guidance for SOs
Research Collaboration & Knowledge Exchange

• Imagery
  • Research Group – exploring papers on CNN architecture, class imbalance, explainable AI
  • Study Group – building core skills through courses and discussion
  • Platform for sharing use cases + discussion of best practice

• Text Classification
  • Platform for sharing use cases + discussion of best practice
Research Collaboration & Knowledge Exchange

• Quality of Training Data
  • Focus on understudied area
  • Identified sources of error that affect model quality

• Model Retraining
  • Examined key concepts of drifts (model, concept, data)
  • Identified factors that enable monitoring and retraining
Research collaboration & knowledge exchange

• Infrastructure Group
  • Exploring supporting environment and infrastructure for scaling up ML projects
  • Discussion of organisation experience on cross-cutting issues such as:
    • Linking ML processes with traditional production processes
    • Generic patterns for ML deployment & servers
    • Roles and capabilities
Monthly Forum

- Main meeting point for the community
- Expert presentations from statistical organisations and academia
- 13 presentations and 7 meetings
- C. 100 members attend each meeting
Sprint @ ONS, July

• 21 members from 14 different organisations
• Model retraining, quality of training data and web scraping data
• Networking with other international data science groups and national statistical organisations
Capacity Building

• Coffee and Coding
  • ML Fundamentals & ML Applications Deep Dive
  • ML Foundations for Non-Programmers
  • Git

• ML strategy workshops
  • UN Regional Hubs for Big Data
  • Middle East, Latin America, Indonesia.
Communications

- Website
- Discussion forum
- Conference presentations
- Guides
- Papers
- Youtube channel
- ML Group video (forthcoming)
- Webinar November 30th
Lessons learned & Impact
Impact

• A place to explore and test ideas
• Sharing of tried-and-tested approaches
• Enabling organisations new to ML to accelerate their development
• Addressing common production challenges
• Raising profile of ML among strategic decision-makers
• Building staff skills
What our members say

“I built awareness of different uses for ML applications and how ML applications combine with other tools in the statistical processes.”

“Learning which projects other organisations are successfully doing helps us allocate our limited ML capacity.”

“It motivated my team to increase their ML skills”

“enabled me to access a vast repertoire of experience on the use of ML for the production of official statistics”

“helped me understand ML in the context of official statistics and government data science”

Membership survey, 2021
Lessons learned

1. Fast-changing field

<table>
<thead>
<tr>
<th>ID</th>
<th>Presentation Title</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Use of ML techniques for classification problems related to CPI</td>
<td>TF-IDF, naive bayes, logistic regression, SVC, SGD, Random Forest, XGBoost, LIME</td>
</tr>
<tr>
<td>2</td>
<td>Matching Big Data to Official Statistics Classifications</td>
<td>direct matching, fuzzy matching, TF-IDF, Best Matching 25, Transformer for translation</td>
</tr>
<tr>
<td>3</td>
<td>Triaging Enquiries using Multilingual Transformers Model</td>
<td>Multilingual BERT, XLM-MLM en-fr, XLM-RoBERTa</td>
</tr>
<tr>
<td>4</td>
<td>Codification of firm activity from free text descriptions</td>
<td>Fasttext, Softmax classifier</td>
</tr>
<tr>
<td>5</td>
<td>New model for coding using Deep Learning</td>
<td>Fasttext, Bi-GRU, Softmax classifier</td>
</tr>
<tr>
<td>6</td>
<td>Unsupervised topic modeling and text classification using top2vec and lab2vec</td>
<td>top2vec, lab2vec</td>
</tr>
</tbody>
</table>

ML methods for text classification used in 2019-20

ML methods for text classification used in 2022
Lessons learned

2. More ML use cases for statistical organisations
   • Text classification and imagery analysis continue to be popular use case
   • Use cases outside usual application areas: Predict the respondents to follow up, create a more evidence-based survey frameworks, Triage the multilingual customer inquiries

3. Quality
   • Different importance for different stakeholders and different use case
   • Unique expertise that statistical organisations (e.g., quality control process) can help
Lessons learned

4. Building infrastructure needed for integration

- Model registry
- Model serving
- Data security
- Monitoring and re-training
- Data management
- Standardisation
- Quality control
- Versioning
- Documentation

Using ML in stat offices....

Data Science Campus
Lessons learned

5. High demand for knowledge exchange and capacity building

- Interest in capacity building for machine learning remains high
- Community is dynamic, enthusiastic and increasingly experienced
- Lack of time resource limits deeper engagement
- Quality and range of activities requires dedicated staff resource
- Partnering with other international groups is beneficial
Future plans
Plans for 2023

• Group to close at end of 2022 as ONS and UNECE redeploy resources to other areas
• Exploring more resource-efficient ways to respond to demand
• Discussions with UN Statistics Division
ML Group 2022 Webinar – 30 November

• Session I 1000-1130 CET
  • Applications of machine learning: web scraping data, text classification, imagery data, AIS data

• Session II 1500-1630 CET
  • Statistical production issues: quality of training data, model retraining, IT infrastructure

• Registration open on Eventbrite
  • Go to the link on the Machine Learning 2022 page [here](#)
Discussion

• What progress has ML made in your organisation? What role do you think the UNECE ML Group has played in that?

• What lessons could the ML Group’s development have for other modernisation projects?

• What are the most important aspects of the group’s work that should be continued? What channels can we consider for this?