Machine Learning Group 2022 - Activity Proposals

CODING AND CLASSIFICATION

1. Automatic coding and classification for problems related to CPIs

Vladimir Miranda

IBGE, Brazil

vladimir.miranda@ibge.gov.br

Name and short outline of the activity/activities you are proposing.Please include brief information about your objectives, methods/techniques to be explored, planned outputs, and potential impact to your organisation or the international ML community.If you are proposing multiple activities please provide a separate outline for each one below.

Our goal is to keep developing the study of classification techniques initiated in the 2021 round. Among different things to explore we aim to improve the results previously developed by combining different data sets and extend the study to other CPI groups beyond food and beverages. We intend to explore traditional ML such as logistic regression, random forests, XGboost in combination with preprocessing techniques. According to time and resources available we would also like to get expertise and test more sophisticated approaches such as those based on word embeddings models. As an output we would like to evaluate what which models perform best for products classification tasks related to CPI problems such as the use of big data sources for CPI purposes. This study can have applications in other areas beyound the CPI division, for instance, for to the classification of products for household budgets surveys. The international community can benefit from this work as a case study in the field and with the material produced as a by-product of the study.

Organisation

Q6. For each activity, please briefly describe how you would plan to organise and run it. Explain how you would involve other group members in your activity (e.g. communicating key results, requesting feedback, assigning them specific research tasks etc) and state what kind of contribution you would expect from members who want to participate.

Other group members could be involved by discussion of related themes, presentation of key results to disseminate the work and collect feedback.

About You

2. NLP models, transfer learning, text classification

Romana Gwizdała

Statistics Poland

Romana.Gwizdala@gmail.com

About Your Activity

Q4. Select the type of activity you are proposing for our 2022 programme (if proposing more than one activity, please select multiple options):

Knowledge Exchange: An activity whose primary aim is to share knowledge and experience on topics of common interest (e.g. a study group on deep learning models for text classification, or a one-off discussion panel or expert talk on input privacy preservation)

Research development: Presenting your own ML project to other members for feedback and advice and to discuss common issues (e.g., receiving feedback on an ML experiment on model-assisted estimation for survey micro-level data you/ your organisation has conducted)

Name and short outline of the activity/activities you are proposing.Please include brief information about your objectives, methods/techniques to be explored, planned outputs, and potential impact to your organisation or the international ML community.If you are proposing multiple activities please provide a separate outline for each one below.

Smart statistical codes classification - Developing NLP models for text classification based on Roberta architecture to provide tools for easy classification based on written description. There are multiple code systems for different areas of business, finance, healthcare etc that can be difficult to navigate. For example EU uses over 600 NACE codes to describe economic activities and owner of new business must provide correct codes for registry. ML based classification tool for Polish language would benefit not only public statistic office but also many bussiness owners.

Organisation

Q6. For each activity, please briefly describe how you would plan to organise and run it. Explain how you would involve other group members in your activity (e.g. communicating key results, requesting feedback, assigning them specific research tasks etc) and state what kind of contribution you would expect from members who want to participate.

Presenting works and results, asking for ideas and feedback. Would love to be pointed toward educational materials and getting any help from ML experts.

3. Coding and Classification for Household surveys

Jael Pérez Sánchez

INEGI, Mexico

jael.perez@inegi.org.mx

About Your Activity

Q4. Select the type of activity you are proposing for our 2022 programme (if proposing more than one activity, please select multiple options):

Knowledge Exchange: An activity whose primary aim is to share knowledge and experience on topics of common interest (e.g. a study group on deep learning models for text classification, or a one-off discussion panel or expert talk on input privacy preservation)

Your proposal

Q5. Short description of the machine learning theme(s) you are interested in exploring in your activity (e.g., coding and classification, editing and imputation, IT infrastructure, model retraining)

coding and classification

Name and short outline of the activity/activities you are proposing.Please include brief information about your objectives, methods/techniques to be explored, planned outputs, and potential impact to your organisation or the international ML community.If you are proposing multiple activities please provide a separate outline for each one below.

During 2020 and 2021 we have explored some ML techniques for coding household surveys, in 2022 we have proposed to bring these algorithms to a productive level in our coding systems.

Organisation

Q6. For each activity, please briefly describe how you would plan to organise and run it. Explain how you would involve other group members in your activity (e.g. communicating key results, requesting feedback, assigning them specific research tasks etc) and state what kind of contribution you would expect from members who want to participate.

For these tasks, at INEGI we are organized into three different groups, the thematic area, the development area and the research area. The thematic area is in charge of the conceptualization of

Q6. For each activity, please briefly describe how you would plan to organise and run it. Explain how you would involve other group members in your activity (e.g. communicating key results, requesting feedback, assigning them specific research tasks etc) and state what kind of contribution you would expect from members who want to participate.

the problems and the execution, the research area of the algorithm development, while the development ones are in charge of the implementation in the different systems.

MODELLING

4. Building a machine learning model to identify spatial clusters within shipping vessel locations data, representing relevant areas within and around shipping ports globally.

Joseph Crispell
ONS-FCDO Data Science Hub
Joseph.Crispell@ons.gov.uk

About Your Activity

Q4. Select the type of activity you are proposing for our 2022 programme (if proposing more than one activity, please select multiple options):

Research collaboration on Data Sets: Working on a common data set with a primary aim to deliver the ML model together with others (e.g., exploring how ML can be used for global AIS data set and developing a ML model with those in other organisations)

Your proposal

Q5. Short description of the machine learning theme(s) you are interested in exploring in your activity (e.g., coding and classification, editing and imputation, IT infrastructure, model retraining)

Building a machine learning model to identify spatial clusters within shipping vessel locations data, representing relevant areas within and around shipping ports globally. The model would be trained based on manually drawn polygons representing known areas for a selection of ports around the world, available through the Automatic Identification Systems (AIS) task team. The output from the model would be predicted polygons representing areas for ports around the world. Note that polygons could be drawn as a post-prediction step based on model outputs.

Name and short outline of the activity/activities you are proposing.Please include brief information about your objectives, methods/techniques to be explored, planned outputs, and potential impact to your organisation or the international ML community.If you are proposing multiple activities please provide a separate outline for each one below.

Activity 1: identify port areas Point locations are available for shipping ports globally (such as through the World Port Index) but these locations don't provide information about the area of the port. To identify key shipping indicators (important for measuring national/international trade and economic indicators) the area of the port encompassing all port activity is required. The area of the port is useful as it allows us to more accurately identify visits to a port (by including all active areas) and the amount of time vessels spend in a port (giving us insights into port efficiency). Usually, port areas are manually defined based on shipping vessel locations data available through the United Nations Global Platform AIS locations service (UNGP) and free tools like the Marine Traffic dashboard or procured through proprietary tools like sea-analytics. Activity 1 aims to use vessel locations data, available through the UNGP, in combination with port areas defined by teams working in the AIS task team to develop a machine learning model that can predict port areas for ports globally. These automatically identified port areas would be defined as polygons and stored within a public port areas database that could be accessed through, and maintained by, the UNGP. These automated polygons could then be improved based on on-the-ground expertise as well as regularly updated using more recent vessel locations data. The polygons themselves don't need to be the output of the machine learning approach, these could be identified based on some post-processing steps (for example drawing a convex hull around grid cells (or hexagons) predicted to be a part of a port by the model). Activity 2: identify berth areas Identifying port infrastructure within port areas is crucial to disaggregating trade and economic indicators and understanding port structure and efficiency. With a defined port area, it is possible to estimate berth areas based on where vessels are reporting as being moored. Current approaches have used simple binning based methods to draw polygons around spatial bins (or grid cells) where there are peaks in the density of vessels reporting as being moored. While preliminary, these methods have been shown to be fairly successful in identifying berths within ports in East Africa. Activity 2 aims to extend existing density based berth detection approaches to use machine learning methods to identify berth areas within the port areas identified by Activity 1. Automatically identified berth areas would be defined as polygons, mapped to a particular port area and stored within the port areas database. Activity 3: identify waiting areas Exploration of the vessel locations data has found that vessels will often wait outside a port. When understanding the efficiency of ports around the world and, therefore, potential trade bottlenecks, it is important to estimate the amount of time vessels spend waiting both within the port area and outside it. Activity 3, based on shipping vessel locations and trajectories, would aim to identify areas where vessels are likely to wait outside of a port. These automated waiting areas would similarly be stored as polygons within the port areas database.

Organisation

Q6. For each activity, please briefly describe how you would plan to organise and run it. Explain how you would involve other group members in your activity (e.g. communicating key results, requesting feedback, assigning them specific research tasks etc) and state what kind of contribution you would expect from members who want to participate.

Ideally the activities associated with the current project proposal would be organised and coordinated through the AIS task team. Groups within the AIS task team would be free to participate, sharing their manual port area mappings and describing any approaches they are exploring to identify port areas. Ideally, drawing on the expertise and experience from the AIS task team would mean there would be data available for multiple ports around the world to inform the machine learning model building, which is important as ports come in all shapes and sizes and the model will need to be robust to that. A port areas team, including members of the AIS task team, FCDO data science hub, and UNECE ML group, would be identified based on those who are keen, and have the capacity, to contribute. This port areas team would refine the current activity proposal into epics, user stories, and tasks, which they can work on together. The port areas team would regularly check in with the AIS task team (through monthly meetings) to report on progress and check in on next steps. For success, the port areas team group members will need to include: UNGP users who are interested in evaluating outputs Existing UNGP users who can share port areas they are currently using in their own analyses AIS task team members involved in UNGP maintenance who can ensure outputs can be maintained going forward and are compatible with existing approaches Machine Learning experts from the UNECE ML group with a background in approaches for spatial clustering Estimating time commitments required is difficult without further refining the project objectives and identifying team members availability but preliminary estimates could be 6-8 months overall with Activity 1 taking 3-4 months including time to get comfortable with AIS vessel locations data and Activities 2 and 3 each taking approximately 1-2 months. To participate in this project a firm grasp of python programming and applying a range of machine learning approaches would be very beneficial given the complexities of the data and UNGP. Other interested AIS Task Team members in the project: Asian Development Bank

5. Model Assisted Estimation

Sorcha O’Callaghan

Central Statistics Office, Ireland

sorcha.ocallaghan@cso.ie

About Your Activity

Q4. Select the type of activity you are proposing for our 2022 programme (if proposing more than one activity, please select multiple options):

Research development: Presenting your own ML project to other members for feedback and advice and to discuss common issues (e.g., receiving feedback on an ML experiment on model-assisted estimation for survey micro-level data you/ your organisation has conducted)

Your proposal

Name and short outline of the activity/activities you are proposing.Please include brief information about your objectives, methods/techniques to be explored, planned outputs, and potential impact to your organisation or the international ML community.If you are proposing multiple activities please provide a separate outline for each one below.

Objectives: Use model assisted estimation to get unbiased estimates of population and sub-population totals for output variables in business statistics surveys using administrative data.
Methods/Techniques: We have trained several ML models (Support Vector Regression, KNN, Random Forest) to predict variables based on a small number of auxiliary variables that are available from admin data for all enterprises. Models are trained using data for survey respondents. Using the model assisted estimator as described here https://projecteuclid.org/journals/statistical-science/volume-32/issue-2/Model-Assisted-Survey-Estimation-with-Modern-Prediction-Techniques/10.1214/16-STS589.ful population totals are estimated. Planned Outputs: Timelier estimates of business statistics outputs. Potential impact to your organisation or the international ML community: Greater use of admin data, timelier outputs.

Organisation

Q6. For each activity, please briefly describe how you would plan to organise and run it. Explain how you would involve other group members in your activity (e.g. communicating key results, requesting feedback, assigning them specific research tasks etc) and state what kind of contribution you would expect from members who want to participate.

Communicate key results via presentation at quarterly update meeting and via written status report. Would specifically request feedback from other group members on how this MAE approach helps incorporate admin data into surveys were response rates have been negatively impacted by COVID pandemic.

6. Spatial Analysis

Shaní Alvarez Hernández

UNODC

shani.alvarezhernandez@un.org

Your proposal

Q5. Short description of the machine learning theme(s) you are interested in exploring in your activity (e.g., coding and classification, editing and imputation, IT infrastructure, model retraining)

Classification, reinforcement learning, ML for cybersecurity, ML for spatial analysis

Name and short outline of the activity/activities you are proposing.Please include brief information about your objectives, methods/techniques to be explored, planned outputs, and potential impact to your organisation or the international ML community.If you are proposing multiple activities please provide a separate outline for each one below.

An exploration of machine learning methods for the spatial analysis of crime data. The goal of this activity would be to review existing literature in order to create a comprehensive body of knowledge that would allow statisticians and GIS experts to approach spatial analysis of crime data with ML tools. As this is a exploratory study, quantitative and qualitative methods could be used to analyze census, geographic, survey, etc, that could improve crime analysis. As for ML learning algorithms, Deep NN, supervised ensemble models, and others could be explored as a first approach. The output of this activity (or series of activities) would be to create a wiki or introductory report for all those who wish to start in ML and combine it with geographic data to produce spatial analyses. It has become increasingly important to be able to add an extra dimension, space, to the study of crime data in oder to enrich the usual analyses that relie on the creation of indexes and summary statistics. This could provide a starting point upon which to integrate this kind of analysis in more research projects.

Organisation

Q6. For each activity, please briefly describe how you would plan to organise and run it. Explain how you would involve other group members in your activity (e.g. communicating key results, requesting feedback, assigning them specific research tasks etc) and state what kind of contribution you would expect from members who want to participate.

By establishing weekly and monthly goals, and presenting the work in progress publicly to other members. Anyone who wishes to participate could provide feedback and volunteer to continue in some specific task or add one of their own. I would mainly expect participants to take the initiative since everyone has different schedules and time zones. However, everything should work toward the weekly or monthly goals established beforehand.

7. Species Distribution Modelling

Sylvie Clappe

Central Statistics Office, Ireland

sylvie.clappe@cso.ie

About Your Activity

Q4. Select the type of activity you are proposing for our 2022 programme (if proposing more than one activity, please select multiple options):

Knowledge Exchange: An activity whose primary aim is to share knowledge and experience on topics of common interest (e.g. a study group on deep learning models for text classification, or a one-off discussion panel or expert talk on input privacy preservation)

Name and short outline of the activity/activities you are proposing.Please include brief information about your objectives, methods/techniques to be explored, planned outputs, and potential impact to your organisation or the international ML community.If you are proposing multiple activities please provide a separate outline for each one below.

Name: Species modelling Outline and objective: Define new methods to model species distributions and compile species richness and biodiversity maps. Methods: Occupancy modelling, species distribution modelling, community modelling could be a start. ML could be implemented at different stage of this project: (i) the conception of the distribution maps per species using random forest algorithm or other more suitable ML methods; (ii) the design of an algorithm able to gather easily and rapidly all the species distribution maps to compile species richness and biodiversity maps. By biodiversity maps, we mean the compilation of maps of biodiversity indicators such as Shannon Index. Outputs: Species distribution maps, species richness maps, maps of biodiversity indicators. Impact: Although the topic is not new in academia, it is new for the ML group and a new area for official statistics at international level. Some work has been done by UNEP-WCMC and Netherlands recently. The development of such methodology could be highly relevant in the field of Ecosystem Accounting whose legal module is being drafted by Eurostat.

Organisation

Q6. For each activity, please briefly describe how you would plan to organise and run it. Explain how you would involve other group members in your activity (e.g. communicating key results, requesting feedback, assigning them specific research tasks etc) and state what kind of contribution you would expect from members who want to participate.

With this proposal, we are looking for experience from other countries. Ideally that would be a teamwork where everyone could be equally involved. It would go from involvement in the design of he method to only feedback depending on who wants to be involved and of the degree of the involvement.

Imagery

8. Satellite Imagery and ML for identifying mixed crops

Therese Uwimana
National Institute of Statistics of Rwanda
therese.uwimana@statistics.gov.rw

About Your Activity

Q4. Select the type of activity you are proposing for our 2022 programme (if proposing more than one activity, please select multiple options):

Research collaboration on Topics: Working on a common topic with other group members with a primary aim of delivering the output together (e.g., investigating ways to re-train ML models and writing a report on good practice with those in other organisations)

Your proposal

Name and short outline of the activity/activities you are proposing.Please include brief information about your objectives, methods/techniques to be explored, planned outputs, and potential impact to your organisation or the international ML community.If you are proposing multiple activities please provide a separate outline for each one below.

PROJECT TITLE: CROPS CLASSIFICATION AND CORRESPONDING CULTIVATED AREA COMPUTATION IN MIXED CROPS FARMS USING SATELLITE IMAGES AND MACHINE LEARNING KEYWORDS OF THE PROJECT: Mixed Crops and Machine Learning PROBLEM STATEMENT The National Institute of Statistics of Rwanda (NISR) conducts Seasonal Agricultural Survey since 2013 in order to provide timely, accurate, reliable and comprehensive agricultural statistics in Rwanda in terms of land use, crop production and livestock to monitor current agricultural and food supply conditions and to facilitate evidence based on decision making for the development of agriculture sector. However, it has been a challenge to automatically identify and quantify individual crops in a multi crops farms to estimate production of major crops (MT), Yield in Kg/ha, cultivated area. Besides, the existing methods of crops identification is done manually. Consequently, NISR proposes a research project on analyzing individual crops in mixed crops farms using Machine Learning techniques on the existing and new high resolution images datasets. OBJECTIVES The objectives of this project are as follow: • Using Big data source for high resolution images to produce Agriculture statistics • Leveraging Machine Learning techniques for Agriculture statistics • NISR Capacity building in Image processing Techniques • Strengthening Research collaboration between National Institute of Statistics of Rwanda (NISR) and ONS- ML team • Expanding the Image Processing and Big data research knowledge to Africa since NISR is the biggest hub for Data Science METHODS TO EXPLORE There are different methods to solve this issue. We intend exploring different images processing techniques with Computer Vision state of art by using Python and/or R packages. PLANNED OUTPUTS • Crops classification in mixed crops farms • Estimation of Individual crops quantity • Individual crops cultivated area computation • Automating individual Crops' statistics

Name and short outline of the activity/activities you are proposing.Please include brief information about your objectives, methods/techniques to be explored, planned outputs, and potential impact to your organisation or the international ML community.If you are proposing multiple activities please provide a separate outline for each one below.

production • Capacity building of NISR employees CONTRIBUTION OF THIS PROJECT TO NISR AND INTERNATIONAL ML COMMUNITY • Sharing knowledge within ML community • Introduction of new methodology for calculating individual crops statistics in mixed crops farms. • Promoting the Regional Hub for Big Data for Africa

Organisation

Q6. For each activity, please briefly describe how you would plan to organise and run it. Explain how you would involve other group members in your activity (e.g. communicating key results, requesting feedback, assigning them specific research tasks etc) and state what kind of contribution you would expect from members who want to participate.

For effective communication among the group members. Trello application will be used for tasks assignment and for tracking the project progress. In addition, all group members will hold a weekly meeting to check the update of the project. A shared Google folder with project management and project related resources will be used for effective knowledge sharing and project tracking. We expect from members who want to participate, the technical expertise in ML for satellite images processing and among the required skills we can mention the below: • Project management • Python and /or R Programmer • Deep Learning packages for Image Processing • Data Pipeline creation • GIS • Technical Advice • Code Review • Testing the technology Furthermore, the access to higher resolution satellite images of mixed crops farms given that the images would be handy since images we have are noisy due to Rwandan climate type.

9. **Convolutional neural networks for learning target variables and extracting image features from EO,**

Joep Burger

Statistics Netherlands

J.Burger@cbs.nl

About Your Activity

Q4. Select the type of activity you are proposing for our 2022 programme (if proposing more than one activity, please select multiple options):

Research collaboration on Topics: Working on a common topic with other group members with a primary aim of delivering the output together (e.g., investigating ways to re-train ML models and writing a report on good practice with those in other organisations)

Your proposal

Name and short outline of the activity/activities you are proposing.Please include brief information about your objectives, methods/techniques to be explored, planned outputs, and potential impact to your organisation or the international ML community.If you are proposing multiple activities please provide a separate outline for each one below.

Official statistics aim to describe society as accurately, timely and efficiently as possible. Estimates of environmental, social and economic variables are snapshots of a country's status. A literal snapshot of a country's status is taken by earth observation (EO). EO provides images taken by satellites and planes. Images may not directly provide the rich and detailed official statistics, but can be used for at least four purposes: 1) early indicators when images are taken more frequently than register or survey data become available (temporal densification), 2) disaggregation by region of statistics from survey samples that are too small for direct regional estimates (spatial densification), 3) more accurate statistics by automatically extracting features from EO that can be used as auxiliary information in the estimation process, and 4) extracting statistical information on themes related to land use. Computers can be trained to learn target variables and extract image features from EO using convolutional neural networks (CNNs), a class of supervised machine learning algorithms within the family of (deep) neural networks. Last year we have implemented CNNs to estimate poverty from aerial images as part of DeepGeoStat, a two-year EU-funded project on "geospatial official statistics from earth observation data using deep learning". This year we aim to improve the first results and expand the experiments. The ONS-UNECE HLG-MOS Machine Learning Group 2022 provides a great opportunity to collaborate with and learn from the international community to further develop the use of EO for official statistics. Our poverty labels are derived from geotagged household income, which cannot be shared for privacy reasons. However, Statistics Netherlands publishes open data on 1-ha and 25-ha squares that could be used as target variables in a shared project, without having to resort to privacy-preserving techniques. Ideas that could be explored include (but are not limited to): - Building networks that can handle more than the usual three color channels (in contrast to networks pre-trained on ImageNet). Satellite images measure other spectral bands and our aerial images also contain near-infrared, which can be used to derive vegetation indices. - Ordinal regression. Supervised machine learning is often divided into classification for nominal target variables and regression for numerical target variables. Neither is optimally fit for ordinal target variables. - Uncertainty quantification. In addition to quantifying the quality of model predictions on test sets it is equally important to quantify the uncertainty of model predictions. We are interested in exploring some of the options suggested in the literature. - Determining image weights. Learning to predict proportions should take into account the size of the denominator. Annotations should take into account the number of annotators and degree of agreement. How do image weights affect test set performance as proxy for performance on truly unseen data? - Dealing with class imbalance. Large variation in class prevalence can drive model predictions towards the majority class. Balancing the training set or using class weights may improve model performance on a balanced test set but may impair model performance on truly unseen and

Name and short outline of the activity/activities you are proposing.Please include brief information about your objectives, methods/techniques to be explored, planned outputs, and potential impact to your organisation or the international ML community.If you are proposing multiple activities please provide a separate outline for each one below.

imbalanced data. - Choosing the optimal set of images to be annotated using active learning algorithms.

Organisation

Q6. For each activity, please briefly describe how you would plan to organise and run it. Explain how you would involve other group members in your activity (e.g. communicating key results, requesting feedback, assigning them specific research tasks etc) and state what kind of contribution you would expect from members who want to participate.

Ideally, we meet other statisticians familiar with CNNs, decide on a research question and answer this together using a common IT infrastructure and open data

10. Creating an open base platform for process and analysis including ML capabilities

Jakob Engdahl, Statistics Sweden
jakob.engdahl@scb.se

About Your Activity

Q4. Select the type of activity you are proposing for our 2022 programme (if proposing more than one activity, please select multiple options):

Research development: Presenting your own ML project to other members for feedback and advice and to discuss common issues (e.g., receiving feedback on an ML experiment on model-assisted estimation for survey micro-level data you/ your organisation has conducted)

Name and short outline of the activity/activities you are proposing.Please include brief information about your objectives, methods/techniques to be explored, planned outputs, and potential impact to your organisation or the international ML community.If you are proposing multiple activities please provide a separate outline for each one below.

Statistics Sweden just received approval of an innovation initiativ together with Örebro university and AI Sweden with the aim to create an open base platform for process and analysis including machine learning capabilities. The solution will be a set of functionality that provides support both for traditional statistical processing (GSBPM-based) as well as training ML-models.

Organisation

Q6. For each activity, please briefly describe how you would plan to organise and run it. Explain how you would involve other group members in your activity (e.g. communicating key results, requesting feedback, assigning them specific research tasks etc) and state what kind of contribution you would expect from members who want to participate.

We would like to connect through the ML 2022 group to share results and to receive feedback from the community of practitioners and experts. If organisations are interested in trying out the solutions or participate in the development, we are open for discussion regaring forms for this.

11. Model retraining; Using satellite imagery to measure social, economic, and environmental activity

Ayoub Mharzi, IMF

amharzi@imf.org

About Your Activity

Q4. Select the type of activity you are proposing for our 2022 programme (if proposing more than one activity, please select multiple options):

Knowledge Exchange: An activity whose primary aim is to share knowledge and experience on topics of common interest (e.g. a study group on deep learning models for text classification, or a one-off discussion panel or expert talk on input privacy preservation)

Research collaboration on Topics: Working on a common topic with other group members with a primary aim of delivering the output together (e.g., investigating ways to re-train ML models and writing a report on good practice with those in other organisations)

Research collaboration on Data Sets: Working on a common data set with a primary aim to deliver the ML model together with others (e.g., exploring how ML can be used for global AIS data set and developing a ML model with those in other organisations)

Name and short outline of the activity/activities you are proposing.Please include brief information about your objectives, methods/techniques to be explored, planned outputs, and potential impact to your organisation or the international ML community.If you are proposing multiple activities please provide a separate outline for each one below.

1- Model retraining: Having been part of the coding and classification working group over the past two year, developing in the first year a PoC (to predict IMF indicators codes) and working in the second year on rolling out a first working version to production, the natural next step in the pipeline would be monitoring the results delivered by this solution and working on retraining the used model. Accordingly, we would be interested to be part of the Model retraining working group aiming to share our ongoing experience and learn from other organizations that are at the same stage or more advanced in their model retaining for a solution in production; 2- Using satellite imagery to measure social, economic, and environmental activity: As Satellite imagery data is being more and more used to derive new indicators, we are interested in using these data to derive economic, socio-economic, climate change indicators. We lack direct experience to lead this research topic, but we are open to explore and build skills and expertise in collaboration with other agencies with common interests.

Organisation

Q6. For each activity, please briefly describe how you would plan to organise and run it. Explain how you would involve other group members in your activity (e.g. communicating key results, requesting feedback, assigning them specific research tasks etc) and state what kind of contribution you would expect from members who want to participate.

1- Model retraining: We would like to participate in this working group to share our experience with implementing the ML solution developed during the first two years. However, due to resource constraints, we are not in a position to coordinate the work of this group. We would suggest that this working group follows the same model used in this past year: - Recurrent meetings: to track work done and discuss specific sub-topics and potential issues faced by group members. During the first occurrences of these meetings, the group can discuss aspects of interest related to retraining Models. In agreement one or multiple organizations part of this group would lead the work and discussions for different sub-topics (e.g. tracking model drift etc.). This would distribute the responsibly of managing the working group across all members. That being said, and similarly to what has been done during this past year, there should be a designated organization overseeing the overall process scheduling meetings, sharing minutes, defining specific agenda items for each meeting etc. - Potentially drop the quarterly update reports for this group as they might not be relevant and potentially cumbersome. These would also be naturally replaced by different presentations delivered during the recurrent meetings. - Group reports: We also suggest instead of having one report per organization for this working, have one common report across the entire group (similarly to what was done for the “from solution to production” group). This report would cover the different sub-topics agreed and each section would be under the responsibility of the organization(s) leading the discussion and work for said sub-topics. 2- Research using Satellite Imagery: the group should begin by doing a stock-taking of satellite imagery data available, with preference for open-source data. Then, the group should identify 2/3 policy relevant questions to answer with the available data. Examples could be estimates of oil inventories from gas plants; density of cargo ships in major ports; consumption activity around shopping areas; estimates of crop yields; etc. ML techniques should be used to extract relevant features from sequence of images over time.

12. Updating supervised learning models for population and characteristics statistics.

Eleanor Law, ONS, UK
Eleanor.law@ons.gov.uk

About Your Activity

Q4. Select the type of activity you are proposing for our 2022 programme (if proposing more than one activity, please select multiple options):

Research development: Presenting your own ML project to other members for feedback and advice and to discuss common issues (e.g., receiving feedback on an ML experiment on model-assisted estimation for survey micro-level data you/ your organisation has conducted)

Your proposal

Name and short outline of the activity/activities you are proposing.Please include brief information about your objectives, methods/techniques to be explored, planned outputs, and potential impact to your organisation or the international ML community.If you are proposing multiple activities please provide a separate outline for each one below.

"Updating supervised learning models for population and characteristics statistics." We are using various methods to train a classifier to predict a person's correct (census) address, given a list of possible addresses on administrative data. We are constructing similar models to predict characteristics e.g. ethnicity. To be practically useful, it will be necessary to retrain models as administrative data sources evolve and change over time. We are exploring how this could be done using historical census and survey data. We are particularly interested in the exploring the best way to use new and old data, the periodicity of updates, and looking at the relationship between the sample size of labelled data provided by a survey and the quality of the updated model. This work will feed into a recommendation in 2023 for the future of UK population statistics.

Organisation

Q6. For each activity, please briefly describe how you would plan to organise and run it. Explain how you would involve other group members in your activity (e.g. communicating key results, requesting feedback, assigning them specific research tasks etc) and state what kind of contribution you would expect from members who want to participate.

We would regularly present our work to the group and share lessons learned. We would ask group members for feedback and discuss common issues.

13. Model retraining and ongoing quality assessment to address model drift.

Eleanor Law, ONS, UK

Eleanor.law@ons.gov.uk

About Your Activity

Q4. Select the type of activity you are proposing for our 2022 programme (if proposing more than one activity, please select multiple options):

Research collaboration on Topics: Working on a common topic with other group members with a primary aim of delivering the output together (e.g., investigating ways to re-train ML models and writing a report on good practice with those in other organisations)

Your proposal

Name and short outline of the activity/activities you are proposing.Please include brief information about your objectives, methods/techniques to be explored, planned outputs, and potential impact to your organisation or the international ML community.If you are proposing multiple activities please provide a separate outline for each one below.

"Simulation studies of model drift for various applications" Group members bring a successful prototype or productionised ML example from their own organisation. A method for perturbing original training data is proposed for each application, and this can be discussed amongst the group to cover various drift possibilities including concept drift. Openly available packages and MLOps tools will be tested to understand how effective/sensitive/efficient they are in different settings for detection of model drift. Results could be drawn together across diverse examples and it may be possible to conclude which methods or tools are most effective or widely applicable.

Organisation

Q6. For each activity, please briefly describe how you would plan to organise and run it. Explain how you would involve other group members in your activity (e.g. communicating key results, requesting feedback, assigning them specific research tasks etc) and state what kind of contribution you would expect from members who want to participate.

Members could work independently on their ML example problems and come together regularly to agree a harmonised approach (e.g. which tools to explore and suitable metrics to use) so that results may be compared across examples. I would not be able to organise this activity, but I think this would be a good opportunity to collaborate with others who are interested, if I have time to do so. The other activity I proposed would be higher priority for us and it may not be possible to be involved in both.