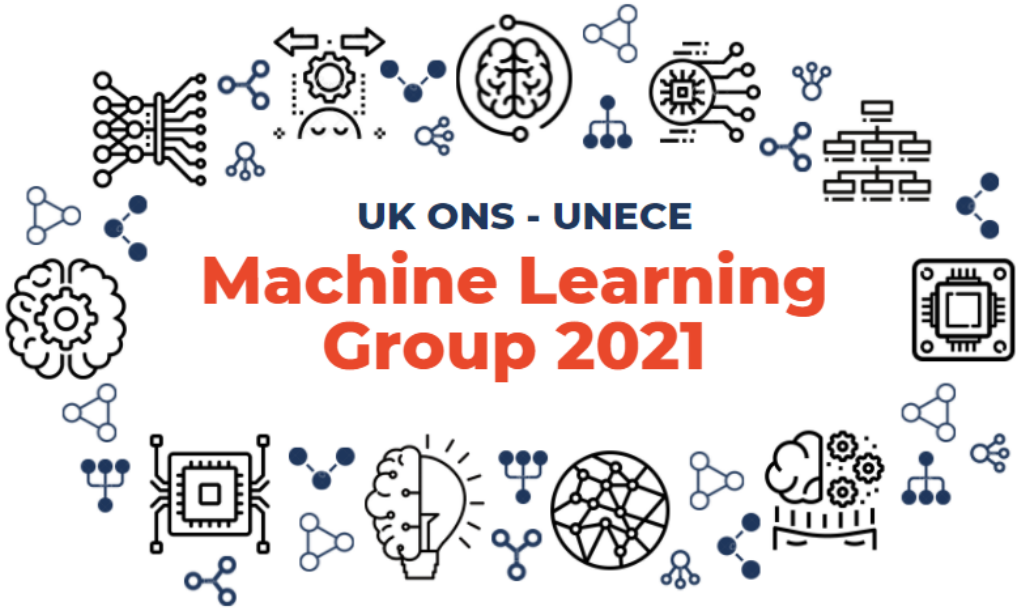


Holistic approaches to evaluating a Machine Learning Model: A case study for automated coding

Jose Jimenez

National Institute of Statistics and Geography

Mexico

INEGI

Agenda

1. Introduction.
2. Use Case Information.
3. Exploring QF4SA Dimensions.
4. Conclusions.

INEGI

1. Introduction

A Quality Framework for Statistical Algorithms

This paper was developed by the 'Quality Aspects' working group of the HLG-MOS Machine Learning project and consisted of: Wesley Yung – Chair (Canada), Siu-Ming Tam (Australia), Bart Buelens (Belgium), Florian Dumpert (Germany), Gabriele Ascari, Fabiana Rocci (Italy), Joep Burger (Netherlands), Hugh Chipman (Acadia University) and InKyung Choi (United Nations Economic Commission for Europe)

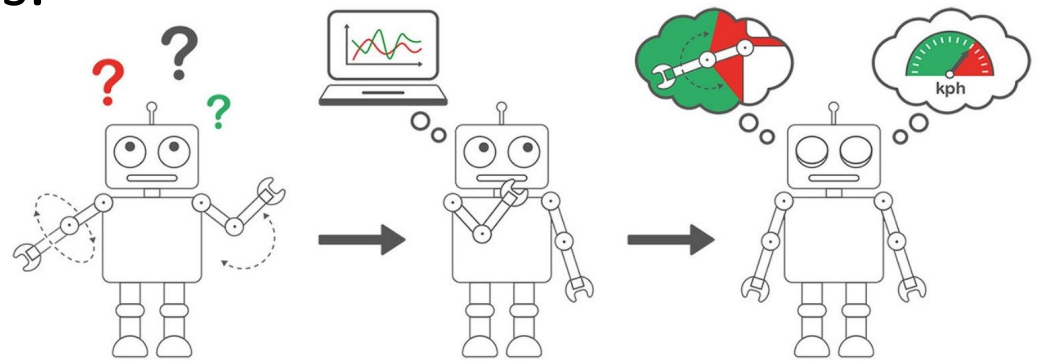
comprises five dimensions:

1) explainability
2) accuracy
3) reproducibility
4) timeliness
5) cost effectiveness.

INEGI

1. Introduction

- **Objective:** To explore the dimensions of QF4SA in a consolidated project to analyse the output of a ML model based on a set of standard metrics and procedures.
- The model was assessed in its initial phase; then a thorough evaluation was performed on it using QF4SA dimensions.



ML ALGORITHMS

2. Use Case Information

Occupation and Economic activity coding using NLP

- Initially, the goal was to leverage Machine Learning models to automate the process of coding regarding such activities.
- The project's initial phase ended in 2020, its use in production is evaluated.
- In 2021, the project moved to a second phase where state of the art techniques were incorporated

Occupation	SINCO	IND_SINCO	Labor	…
Taxi Driver	4586	3	To move people	
Owner	1111	2	Pay, Sell merchandise	

INEGI

3. Exploring QF4SA Dimensions

I. Explainability

- The first approach to show Explainability was to shuffle the values in the columns: col_1, col_2, …col_n, changed to col_n, … col_2, …col_1.
- These changes were made in text and numerical columns used in the model.
- The results for SVM were:
  - Accuracy: 88.37%
  - Time: 82 minutes
- Results show the classifier can have a different output according to input data.

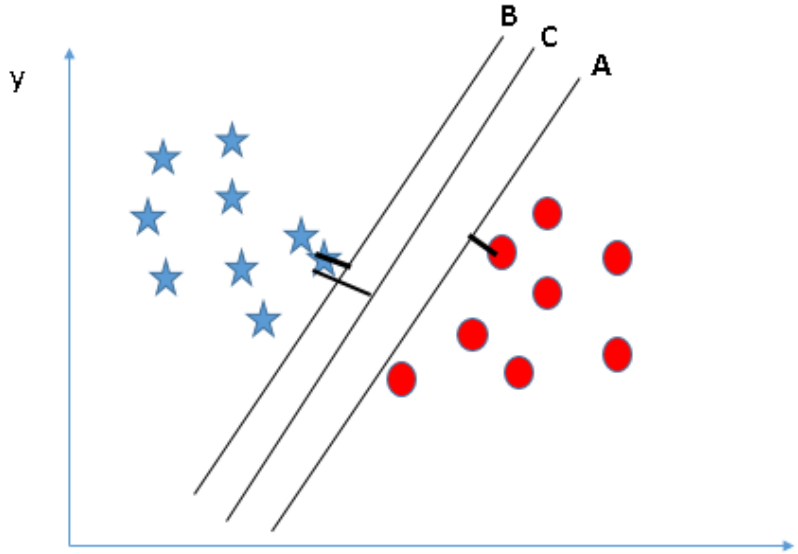
INEGI

UK ONS - UNECE
Machine Learning
Group 2021

3. Exploring QF4SA Dimensions

I. Explainability

- The second approach was to use an adversarial example, by changing the values in the text columns.
- Different words were added as a prefix of the column.
- The results for SVM were:
  - Accuracy: 88.22%
  - Time: 97 minutes
- These results show the classifier can have a different output according to input data. Execution time was higher.



INEGI

3. Exploring QF4SA Dimensions

II. Accuracy

- These ML classifiers were used for classification: Random Forest Classifier, Extra Trees Classifier, Multi Layer Perceptron Classifier, Logistic Regression, Linear SVM
- Linear SVM was the classifier with the highest accuracy: 88.32%
- Several experiments were performed by INEGI colleagues with the goal of increasing the initial accuracy achieved by LSVC
- Some results are shown next

INEGI

3. Exploring QF4SA Dimensions

II. Accuracy - Results

Freq Class	f1_macro	accuracy	recall	precision	time/secs	Model	Dimensions	f1_macro	accuracy	recall	precision	time/secs	Freq Class
>=100	0.7651	0.874	0.7846	0.754	100.26	LSVC	25000	0.7829	**0.8839**	0.7997	0.7718	1202.88	>=100
>=100	0.7638	0.8735	0.7837	0.7522	94.61	LSVC	20000	0.7813	0.8833	0.7986	0.7693	1252.39	>=100
>=100	0.7607	0.8729	0.7818	0.7478	88.04	LSVC	15000	0.7781	0.8823	0.795	0.7667	1173.1	>=100
>=100	0.7569	0.8708	0.7789	0.7437	X-95.97	LSVC	10000	0.7759	0.8807	0.7935	0.7642	1065.43	>=100
>=40	0.7104	0.8706	0.737	0.6962	121.06	LSVC	25000	0.7251	0.8787	0.7506	0.7112	1365.33	>=40
>=40	0.7086	0.8692	0.7357	0.6946	114.46	LSVC	20000	0.7252	0.8786	0.7503	0.711	1444.81	>=40
>=40	0.7083	0.8687	0.7369	0.6929	106.12	LSVC	15000	0.7218	0.878	0.749	0.7064	1410.97	>=40
>=40	0.7026	0.8663	0.7324	0.6866	X-101.01	LSVC	10000	0.7203	0.8765	0.7507	0.7022	1366.65	>=40
>=4	0.6461	0.8671	0.673	0.6398	141.56	LSVC	25000	0.6518	0.8755	0.6744	0.6464	1638.27	>=4
>=4	0.6436	0.867	0.6703	0.6388	147.34	LSVC	20000	0.6517	0.8753	0.6762	0.6432	1551.18	>=4
>=4	0.6426	0.8652	0.6717	0.6367	X-150.07	LSVC	10000	0.6428	0.8748	0.67374	0.6315	X-1397.16	>=4
>=4	0.6426	0.8667	0.6733	0.6351	136.91	LSVC	15000	0.6498	0.8754	0.6756	0.6405	1547.78	>=4

3. Exploring QF4SA Dimensions **II. Dimensionality Reduction**

Freq Class	f1_macro	accuracy	recall	precision	time/secs	MODEL	Dimensions
>=100	0.7506	0.8715	0.7425	0.784	4.2	ExtraT	10000
>=100	0.7446	0.8671	0.7377	0.7738	4.2	ExtraT	15000
>=100	0.7351	0.8602	0.7294	0.7623	4.38	ExtraT	20000
>=100	0.7317	0.858	0.7261	0.7553	4.43	ExtraT	25000
>=40	0.6909	0.8674	0.6868	0.7279	5.51	ExtraT	10000
>=40	0.6812	0.8618	0.683	0.71	5.42	ExtraT	15000
>=40	0.6681	0.8573	0.6717	0.6996	5.74	ExtraT	20000
>=40	0.6614	0.8522	0.6648	0.69	5.72	ExtraT	25000
>=4	0.6145	0.8609	0.6076	0.6643	X-52.89	ExtraT	10000
>=4	0.6054	0.8566	0.5994	0.648	7.76	ExtraT	15000
>=4	0.5895	0.8508	0.5852	0.6348	7.8	ExtraT	20000
>=4	0.5837	0.8457	0.5817	0.6211	8.06	ExtraT	25000

LDA + Aux Var

Freq Class	f1_macro	accuracy	recall	precision	time/secs	MODEL	Dimensions
>=100	0.7773	0.8814	0.7951	0.766	18846.42	LSVC	25000
>=100	0.7765	0.8808	0.7956	0.7635	17224.23	LSVC	20000
>=100	0.7758	0.8802	0.795	0.763	13427.43	LSVC	15000
>=100	0.7689	0.8772	0.7899	0.7547	X-6554.18	LSVC	10000
>=40	0.721	0.8767	0.749	0.704	25191.93	LSVC	25000
>=40	0.7192	0.8771	0.7481	0.7024	20345.12	LSVC	20000
>=40	0.7158	0.876	0.7428	0.6989	15532.49	LSVC	15000
>=40	0.7122	0.8733	0.7431	0.6933	7245.52	LSVC	10000
>=4	0.6505	0.8745	0.6758	0.6436	23114.52	LSVC	20000
>=4	0.6489	0.875	0.6746	0.6403	24382.86	LSVC	25000
>=4	0.6466	0.8744	0.6743	0.6377	17573.87	LSVC	15000
>=4	0.6338	0.8725	0.6676	0.6208	X-8295.43	LSVC	10000

PCA + Aux Var

3. Exploring QF4SA Dimensions

II. Accuracy – Class Balancing

Freq Class	Class Balance					MODEL	Dimensions	Class Balance + Var Aux					Freq Class
	f1_macro	accuracy	recall	precision	time/secs			f1_macro	accuracy	recall	precision	time/secs	
>=100	0.7559	0.7623	0.7623	0.7638	4.11	LSVC	20000	0.7739	0.7816	0.7776	0.7782	51.64	>=100
>=100	0.7484	0.7539	0.7539	0.7511	3.61	LSVC	15000	0.7713	0.7797	0.7797	0.7733	51.23	>=100
>=100	0.7475	0.7524	0.7524	0.7605	3.6947	LSVC	10000	0.7692	0.7772	0.7764	0.7716	46.7	>=100
>=100	0.7357	0.741	0.741	0.7404	4.33	LSVC	25000	0.7656	0.7722	0.7722	0.7694	51.26	>=100
>=40	0.6997	0.7023	0.7023	0.7237	193.07	MLPC	10000	0.6959	0.7182	0.7168	0.7101	25.27	>=40
>=40	0.6905	0.7043	0.7043	0.7092	2.28	LSVC	20000	0.6771	0.6924	0.7035	0.6918	25.87	>=40
>=40	0.6891	0.6954	0.6954	0.7094	2.41	LSVC	25000	0.6718	0.6815	0.6888	0.6905	357.23	>=40
>=40	0.6814	0.6934	0.6934	0.6979	2.23	LSVC	15000	0.6685	0.6845	0.6912	0.6858	33.11	>=40

INEGI

UK ONS - UNECE
Machine Learning Group 2021

3. Exploring QF4SA Dimensions

III. Reproducibility - Considerations

- Currently, many research studies are difficult to reproduce independently
- Sometimes data are partly available or not available at all
- NSOs have data that can be shared externally, however, there also exists internal not open data
- Methods Reproducibility could be an alternative for NSOs when sharing their experience in developing ML models

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider,

INEGI

3. Exploring QF4SA Dimensions

III. Reproducibility - Guidelines

- Commonly, data scientists have a different logic as well as diverse technical skills, therefore, documentation in-code is highly advisable
- This documentation practices are also a part of maintainability
- Documenting details of how the model was trained
- It is also recommended to perform control versioning in both training data and feature generation
- Provide details of the software used to construct the ML model such as versioning and packages used

INEGI

3. Exploring QF4SA Dimensions

III. Reproducibility at INEGI

- Coding occupation activity involved comparing SVM (best clf) with other ML methods
- Personal experience around this coding activity project
- Inferential reproducibility using a deep learning approach
- There is currently interest in using a similar approach

  on our national survey of occupation and employment

- As such, this project is intended to be the first transfer learning task at INEGI

3. Exploring QF4SA Dimensions

IV. Timeliness

- The length of time between the reference period and the availability of information
- Data cleansing: 2 Weeks approximately; **Remember** the time spent in this task
- Informatics infrastructure: INEGI Infrastructure; Considerations, On-premise vs. Cloud
- Preparation of training data: First ± 1 week, then it may become a repetitive task
- Evaluation of data quality: Benchmarks, Consensus, Review
- Scalability of the approach: No tests have been performed to evaluate datasets from a different dimension (thus, evaluate scalability).

INEGI

3. Exploring QF4SA Dimensions

V. Cost Effectiveness

Potential additional fixed and ongoing costs for machine learning adoption

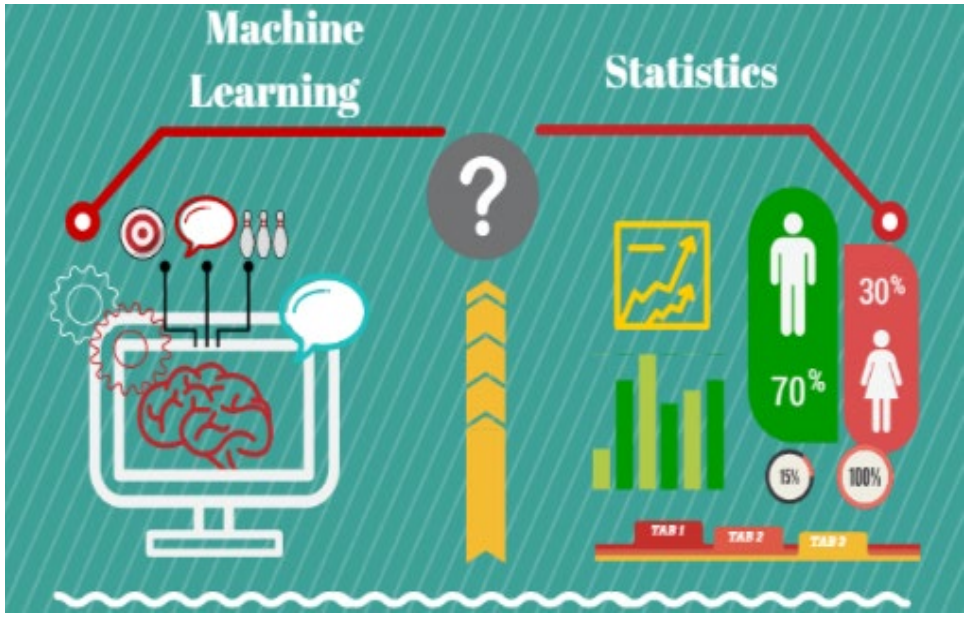
Cost component	Type	Purpose	Comparison
IT infrastructure	Fixed	Acquiring necessary hardware and software	- No dedicated unit, but is part of our sandbox. - Less computers dedicated to this task
Cloud storage	Ongoing	Acquiring necessary cloud storage space	- *Cloud constraints*
IT maintenance	Ongoing	Maintaining IT infrastructure	- Lot cheaper than maintaining several computers - Code maintenance should be considered
Initial staff training	Fixed	Training current staff on ML; may include hiring new staff	- Cheaper than training new human coders - LCiD Team
Ongoing staff training	Ongoing	Keeping staff up to date withnew ML developments	- This is a cost to consider
Data acquisition	Fixed/ ongoing	Acquiring and processing new data sources	- Because of the relevance of the survey, normal data will be estimated, new sources might be integrated
Quality assurance	Ongoing	Conducting quality assurance and control	- This could be an internal or external control

INEGI

UK ONS - UNECE
Machine Learning Group 2021

4. Conclusions

- More ML projects should be evaluated using QF4SA, not only those from NSOs
- The framework must be periodically revised to increase the dimensions, or to improve current ones
- Deep Learning models should be considered in the dimensions and integrated in the evaluation.
- Output from ML models can be analyzed using the framework and compared vs their metrics



INEGI

Thank you for your time

INEGI