The Use of Artificial Intelligence and Machine Learning in DOS

By Malcolm Cai, Jun Wen Tay and Nicholas Chin
Digital Services and Transformation Division
Singapore Department of Statistics

Introduction

The world is in a transformative era fuelled by the rapid development of new Artificial Intelligence (AI) and Machine Learning (ML) technologies. They have the ability to revolutionise operations, products, and services in many domains, such as statistics, from smart automation and data capabilities, to hyper-personalised customer experiences. To thrive in an AI-driven world, organisations have to embrace change and transform the way they operate. They would also need to be aware of the growing need for better governance to tackle ethical, privacy and accountability issues associated with AI deployment.

The Singapore Department of Statistics (DOS) has embarked on a transformation journey in the past years to deploy AI/ ML technologies in statistical processes. This article discusses three key areas critical to AI/ ML implementation – business processes, human capital, and governance framework.

Tapping Ever-Evolving Technologies Requires a Focused and Agile Approach

DOS adopted a focused and agile approach to embed AI/ ML and automation technologies in business processes. The Digital Transformation Unit was formed to work closely with various divisions in DOS and acts as a coordinator to identify synergies between existing and potential AI/ ML projects across the data value chain – from data collection to analysis and dissemination. Another role of the Unit was to identify and prioritise projects with well-defined use cases[1] to operate within resource capacities. AI was only utilised when it is best suited in meeting user needs. Once the value of the AI/ ML method is established via proof-of-concept[2] experiments, the Unit would then optimise and productionise it for use, roll it out across DOS and offer it to Research and Statistics Units (RSUs), and agencies within the Whole of Government (WOG).

Creating a Vibrant Talent Ecosystem

The quality of manpower is critical for successful AI/ ML adoption. Within DOS, statisticians' expertise in statistical theory and computer programming placed them in a good stead to pick up advanced AI/ ML skillsets. Beyond training courses to upskill staff, the Unit created communities of practices and various channels to facilitate information exchange and ideation, and promote cross-collaboration within DOS, RSUs, and WOG agencies. Secondment of staff to private companies with strong data science and AI/ ML capabilities allowed DOS to glean best practices from the private sector and benchmark against industry standards. Thematic sharing on data science and analytics between RSUs, government agencies and industry leaders have helped to create a conducive exploration environment for innovation, and opened new possibilities in using AI to address business challenges.

Data Stewardship and Governance as a Cornerstone for AI/ ML Adoption

As the National Statistical Office, the advocation of statistical best practices and robust data management by DOS have ensured that quality data are produced within DOS and across the WOG. DOS's experience in data stewardship and governance has laid a strong foundation for AI/ ML adoption. Key principles of responsible and ethical AI, such as transparency, interpretability and accountability, are part of the broader framework of statistical best practices. DOS's track record of data management, privacy, security,

1 Well-defined use cases refer to projects where the problem statement, objectives, and expected outcomes are clearly described. The area where AI/ ML is used should be mentioned and the benefits should be measurable.

2 Proof-of-concept measures the feasibility of a project by gathering evidence to support or through a demonstration.

and accountability maintains public's trust in DOS's data products and services, and contribute towards the building of a digital government through the government data architecture[3] initiative.

DOS's efforts have led to successful outcomes. Over the years, DOS has developed many in-house solutions that optimise statistical operations, ranging from data collection, compilation, analyses, and dissemination. Examples include finding a novel way of updating the Statistical Business Register[4], automating statistical processes such as the classification coding in Census 2020[5], and providing new statistical insights[6].

The following sections elaborate two of DOS's latest AI initiatives - the DOS Intelligent Classification Engine (DICE) to process natural language text into Singapore Standard Classification codes and the ML Toolkit to simplify complexity surrounding ML projects.

DOS Intelligent Classification Engine (DICE)

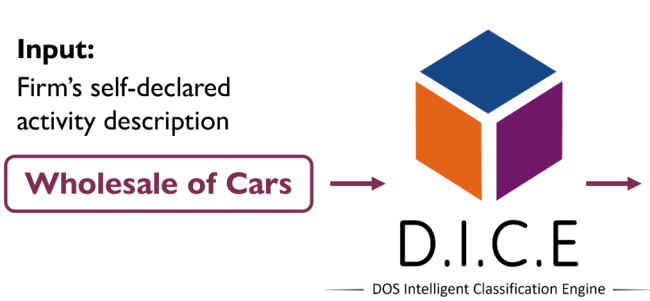
The Singapore Standard Industrial Classification (SSIC) is used to classify firms according to their principal economic activity. As data on the SSIC are widely used by public agencies in many aspects (e.g., compilation of economic statistics, research studies, and policy implementation), it is crucial to ensure the accuracy of SSIC data. However, firms may find it challenging to select the most appropriate SSIC code when registering with the Accounting and Corporate Regulatory Authority (ACRA). Agencies surveying firms may also find it difficult to classify firms into sectors based on activity descriptions of the firms. The effort to ensure the correct classification of firms is significant, involving coordination between agencies, manual verification of SSIC codes, as well as training staff to be well-versed in SSIC knowledge.

Thus, DICE was developed for users (e.g., firm registrants and public officers) to automatically identify the most relevant SSIC codes through ML. DICE takes a text description of a firm's economic activity as an input and recommends suitable SSIC codes[7] as an output, along with a score on how confident the engine is on the predictions (Figure 1).

To train DICE, the team acquired large quantities of data on the descriptions of principal activity and the corresponding SSIC codes from sources such as administrative data from ACRA, survey returns from DOS and RSUs, and official SSIC definitions. However, curating clean training data with the correct SSIC labels is challenging, since a proportion of firms from the raw data sources are expected to have wrong SSIC labels. The time and effort required to manually verify voluminous number of records is impractical. Hence, an in-house solution was developed to automatically detect wrongly labelled data for correction.

In addition, the team tapped on BERT[8] to enhance DICE's capabilities. BERT is a large language model developed by Google and trained on a large corpus of

FIGURE 1 USING DICE TO RECOMMEND SSIC CODES



Prediction	Score
SSIC 46561: Wholesale of motor vehicles except motorcycles and scooters	92%
SSIC 46563: Wholesale of parts and accessories for vehicles	5%
SSIC 47311: Retail sale of motor vehicles except motorcycles and scooters	2%

3 More details on government data architecture are available from the Digital Government Blueprint.

4 DOS (2023), 'Experimental Uses of Machine Learning and New Data Sources in Updating the Statistical Business Register' in Statistics Singapore Newsletter (SSN), Issue 1, 2023.

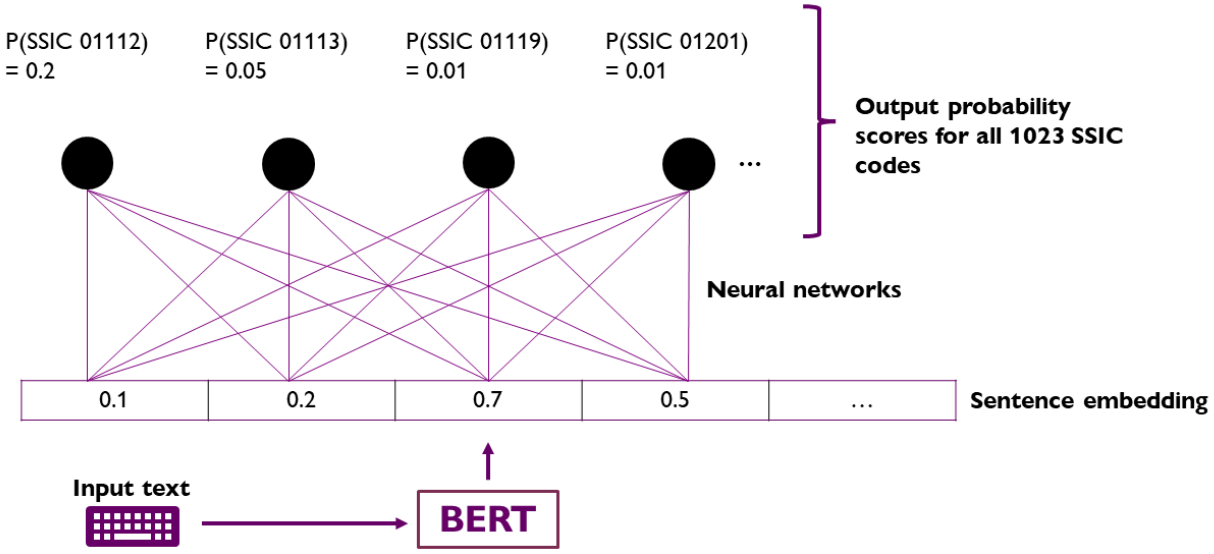
5 DOS (2021), 'Coding of SSOC/ SSIC in Census 2020 using Machine Learning' in Statistics Singapore Newsletter (SSN), Issue 2, 2021.

6 DOS (2022), 'Using Big Data to Profile Singapore's Internet Economy' in Statistics Singapore Newsletter (SSN), Issue 2, 2022.

7 DICE can generate recommendations at various levels of SSIC granularity, e.g., 5-digit and 2-digit.

8 Bidirectional Encoder Representations from Transformers. Read more in: Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018), 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'.

FIGURE 2 SIMPLIFIED ARCHITECTURE OF DICE



text including 2.5 billion words from the English Wikipedia. BERT models utilise sentence embeddings to support various downstream tasks such as predicting SSIC codes, by adding a neural network layer on top of them (Figure 2). The entire network is then fine-tuned by training on the task-specific data (i.e., descriptions of firm activities and corresponding SSIC codes).

DICE has fared well across several performance metrics, like having an accuracy score of around 86%. With DICE, large volumes of data are now more efficiently and accurately classified within a shorter time frame as compared to manual labelling.

The project team continues to enhance DICE by working closely with partner agencies such as the Economic Development Board, Maritime and Port Authority of Singapore, and ACRA, on their use cases to ensure that DICE's functionality aligns with their operational needs. The team is also working with the Ministry of Manpower for DICE to support the Singapore Standard Occupational Classification (SSOC) code prediction. Moving forward, DICE is envisaged to be a WOG productivity tool for Singapore Standard Classification use cases in classifying codes related to expenditure, commodities, education, etc., beyond industry and occupation codes.

Machine Learning Toolkit

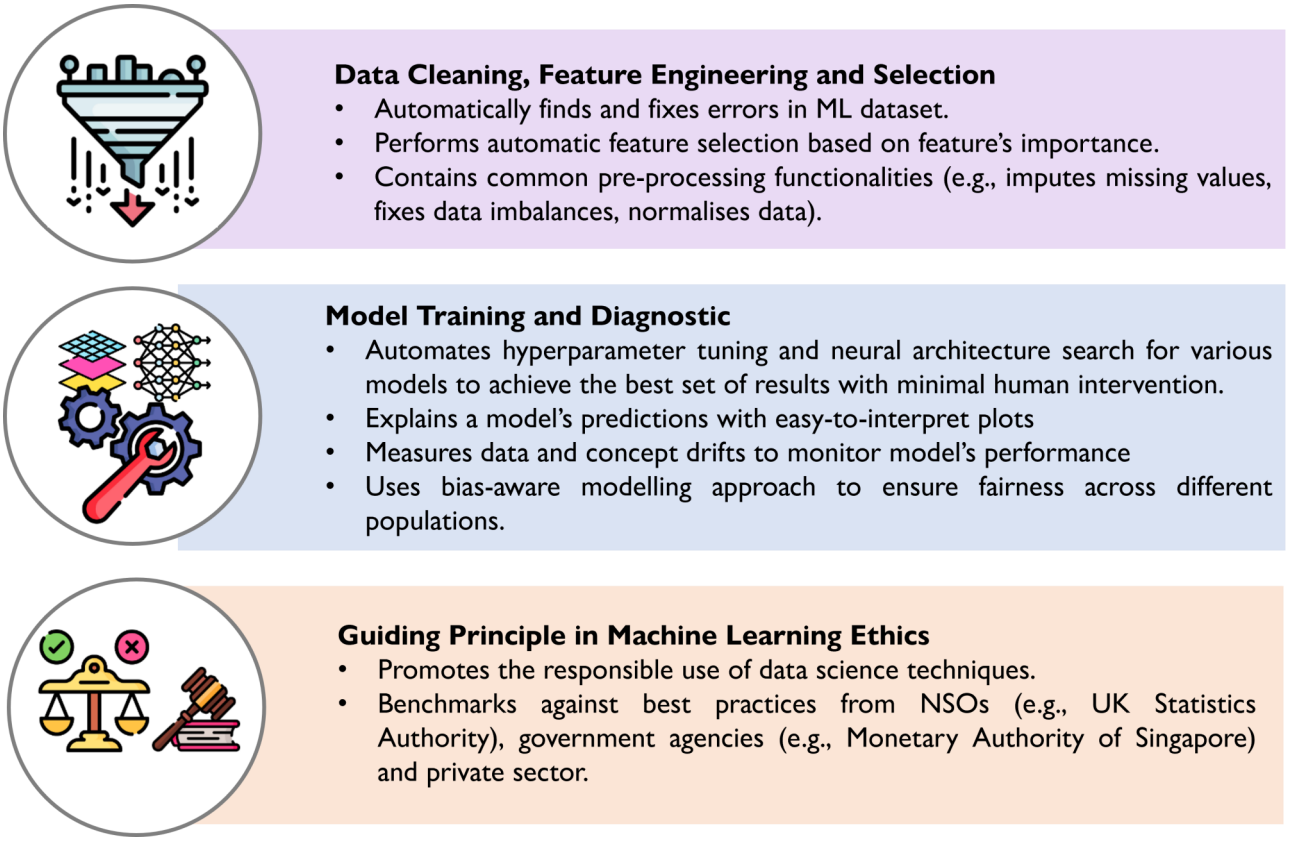
Increasingly, ML is being used in data analysis and business decision-making. However, the complexity of ML models and its steep learning curve make it challenging for those with limited data science background to implement ML projects efficiently in DOS. Furthermore, building ML models is time-consuming and often requires trial-and-error for the model's hyperparameters.

While commercialised automated ML (AutoML) products can offer comprehensive end-to-end ML capabilities, they come at a premium price. With the aim to empower and support DOS statisticians in adopting ML for their work, acquiring commercial products at a large scale would not be cost effective.

As such, the Unit conducted extensive research on various popular open-source AutoML libraries and curated a selection of suitable ones to develop DOS's own toolkit. This helps to lower the barriers to entry for data science exploration and makes it easier for officers to build ML models. Figure 3 details the end-to-end process from data cleaning to model training and diagnostics. It also reinforces best practices in ML ethics such as evaluating fairness of a model and providing transparency in managing ML projects.

Statisticians in DOS can now easily conduct their ML modelling process by answering simple questions for the toolkit to determine the best approach for the data and task. Thereafter, the toolkit will train and build different ML models and provide relevant model information such as performance metric, fairness metrics, and interpretability plots. With these outputs, DOS officers can select the better-performing models

FIGURE 3 FEATURES OF THE MACHINE LEARNING TOOLKIT



to further explore or fine-tune for their use case. This increases efficiency significantly without having to build the ML model from scratch and the officer will only have to focus on a smaller subset of models output from the toolkit.

As ML models are increasingly being used to make decisions that can have significant impact on individuals and the society, it is important to ensure that the model developed is fair. The toolkit is able to compute fairness metrics and generate fairness plots to highlight unfair models, while providing sample codes to utilise techniques such as resampling and reweighting to improve the model's fairness. This ensures that the developed model will not discriminate against certain groups.

Overall, the ML toolkit encourages a hands-on approach to experimentation and to test and refine predictive models and analytical techniques. It helps statisticians streamline ML processes, therefore enhancing DOS's overall capability.

Conclusion

The gradual but fundamental shift in adopting AI/ ML technologies evolves DOS's business processes, upskills manpower, and updates governance framework. It enables DOS to harness AI productively, which in turn enhances the data products and statistical services offered to various user groups.

DOS will continue to invest in AI training for staff, such as developing an internal AI playbook to keep abreast of the latest advancements in both AI and generative AI spaces, as well as practical applications of AI in the statistical and data analytical fields.

Moving forward, other AI initiatives in areas such as data collection (e.g., web-scraping as a new key data source, document intelligence) and data dissemination (e.g., seamless user experience in searching for relevant information and data) are actively explored, as part of DOS's efforts to continually enhance the process of collecting, producing and consuming statistical data.