**Describe the Machine Learning Pipeline**

**Introduction**

This paper serves as a submission for a university assignment and provides a comprehensive outline of the machine learning process. The author will guide the reader through this process in seven distinct stages. In addition to theoretical descriptions, the paper will also explore the application of machine learning in various contexts, with an emphasis on healthcare, reflecting the author's interests. Likewise, the author also wishes to briefly discuss his own perspectives reflecting how simply describing the pipeline is just one aspect of this evolving, complex and novel subject.

**Background**

Machine Learning, hereon abbreviated as per convention to ML, is a subfield of Artificial Intelligence (AI) (ML Cheatsheet, 2017). ML combines complex statistical models with large datasets to create the capacity of a computer to learn and make decisions without the need for this process to be explicitly coded for by a human programmer (ML Cheatsheet, 2017). Although ML has been a tool which has been under development for several decades, more recent innovations in Computer Science and the widespread use of data led to advancements in ML of such magnitude that AI has been described as heralding *“The Fourth Industrial Revolutio*n” (Velarde, 2019). Although this may come across as a bold claim, in the next paragraph we will describe the current and predicted impact of ML and hope to convince the reader that this present surge of interest in AI will not be like previous cycles of “hype” and subsequent disinterest (Gartner, 2023)- but that ML is here to stay as a force of disruption.

In the financial industry for instance, ML excels in processing vast scales of market data for to perform prediction, useful for tasks such as risk analyses, and decision making- similarly useful for tasks such as trading (Goodell et al., 2021). Its efficiencies in these tasks, far surpasses what a human finance worker is capable of, marking ML as a transformative FinTech innovation, with a Bank of England Survey revealing 79% of firms had ML strategies in place. Similarly Natural Language Processing (NLP) technologies, a subset of ML, has highlighted promise to greatly transform scientific research and development- the utilisation of AI for Drug Discovery being an exciting example of this (Paul et al., 2021). NLP’s have given rise to technologies such as Large Language Models (LLM) and voice assistants-technologies set to transform human interaction with technological products (Vajjala et al.,2020). The ability of ML to guide smart robotics and integrate other technologies into networks we call Internet of Things, reflect the potential for ML to transform the manufacturing process into streamlined, autonomous systems (Strielkowski et al., 2023). ML is of particular interest to the author due to its ability to disrupt medical imaging; market research estimated the US market of ML Medical Imaging to be around $1 billion USD in 2023 and set to grow twenty-fold over the next decade (Grand View Research, 2023).

Having outlined how ML is rapidly transforming Science and Industry, we shall navigate the world of ML by outlining seven stages pertinent to the development and deployment of ML technologies. In addition, we shall also be evaluating the relative strengths and weaknesses of different ML models. For those seeking to disrupt their own fields of interests, understanding the steps and missteps in ML execution is imperative for the creation of an effective and scalable ML model.

It is also important to note that not every data science project necessitates the implementation of a pipeline. In scenarios where data scientists are engaged in experimental tasks, such as investigating new model architecture or replicating recent publications, pipelines may not be deemed essential. We hope that this paper provides an outline of why and how the reader ought to contemplate pursuing their own ML project.

**Detailing the ML pipeline**

In this section of the paper we will now move on towards outlining the Machine Learning Pipeline.

To help explain the Machine Learning Pipeline, we will be using the diagram below as shown in Figure 1. Figure 1 demonstrates our recommended workflow and how the training data will undergo seven stages for processing.

A diagram of a company

Description automatically generated

Figure 1. Workflow of Machine Learning (ML) Pipeline. (Khan, 2024)

As shown in Figure 1, our seven stages are: Data Acquisition, Data Cleaning and Preparation, Data Exploration and Understanding, Model Building and Training, Hyperparameter Optimisation, Model Evaluation and Model Deployment.

Let us explore below, in a systematic way, each of these steps in turn.

Step 1: Data Acquisition

The selection of appropriate data sources is paramount in ensuring the accuracy and reliability of research endeavours. Mismatched data and research questions can lead to biased or incomplete results, underscoring the importance of prioritising relevance, reliability and representativeness. Therefore, data acquisition focuses on obtaining data that is not only amenable to cleaning and pre-processing but also serves as a foundation for training machine learning models. Online databanks, such as data.gov, AWS Public Datasets, and Google Cloud Public Datasets, play a pivotal role by offering structured data tailored for quantitative analyses across diverse domains. Simultaneously, scientific literature serves as a rich source for qualitative insights, with platforms like IEEE Xplore, Google Scholar, and the Data Citation Index providing valuable scholarly resources. The author wishes to highlight that for his own interest in liver tumours there are several publicly available datasets generated from competitions held to allow various Deep Learning (itself a subset of ML) compete in identifying the liver and liver tumours from medical images; these datasets are often utilised as benchmark tests by other developers down the line. One of the most notable datasets is the Liver Tumour Segmentation Benchmark 2017 (LiTS17) developed from a competition held in 2017 (Bilic et al., 2023). It must be noted that there are significant concerns over the handling of the sensitive data and that those involved with providing ML products to the public do not discriminate and inflict harm to their end users because their training data may not have been of optimum quality (Velarde, 2019).

Step 2: Data Cleaning and Preparation

After having acquired data, it is essential that we ‘clean’ and prepare it. This process means that we try to get rid of inconsistencies and inaccuracies in our data and then make it suitable and standardised for our model to use.

Some basic steps we can do to make our data clean is through a visual exploration of our data. We can carry out basic steps such as creating box-plots and histograms to identify outliers. Missing data may also be an issue when trying to harness training data. Whilst one option could be to delete parts of our data, so that only full datasets are preserved, this has the disadvantages of reducing our pool of data, losing out on valuable data and potentially introducing bias. Another method would be to employ imputation techniques such as mean imputation. In the domain of medical imaging having complete and consistent data is especially important.

After cleaning, it’s important to have the data prepared into a standard format. This will make it significantly easier for our model to process our data. We can also describe this as ‘normalising’ our data. Normalisation techniques include Z-score normalisation and min-max scaling. With images having scales which can vary widely, such as 8 bit image pixels which can have a scale from 0-255; with normalisation we can convert this to a scale from 0-1, a much easier scale for our model to computationally process.

Step 3: Data Exploration and Understanding

A robust machine-learning pipeline begins with a thorough exploration of the dataset revealing patterns, identifying outliers and assessing data quality. Features will then be selected with regards to the target outcome. Dimensionality reduction techniques, such as Principal Component Analysis (PCA), are employed to manage high-dimensional data effectively. Through these steps, the performance of the model is significantly enhanced, improving its accuracy and interpretability. This process not only reveals underlying patterns within the dataset but also leads to a deeper and more reliable analysis, contributing to advancements in the application field. By integrating these essential data preparation steps, the machine learning model becomes more adept at understanding complex data structures, ultimately supporting more informed decision-making.

Step 4: Model Building and Training

In the ML pipeline, it is important to tailor the model with regards to the data that we have and the problems that we are trying to deal with. Different models cater to varied data types and learning tasks. Linear models may be more straightforward in linear relationships say BMI and the risk of developing Type 2 Diabetes; in this case we may wish to use a linear regression model. An example of a non-linear relationship could be the exponential growth of bacteria, in which case we may want to use a Random Forest Model. Strategic selection of the most appropriate model becomes crucial, aligning with the nuanced characteristics of the dataset and problem at hand.

Our linear regression model is an example of supervised learning as it is a process which relies on explicit labels. Conversely unsupervised learning where labels are not explicitly assigned may be useful in tasks where we wish to uncover new and new knowledge about things which we may not even be aware of. For instance, we may use unsupervised learning to discover new insights about consumer spending habits-without necessarily knowing what classes of information we are seeking to link together. Semi-supervised learning is a hybrid approach which we may use when we have a large amount of unlabelled data and a small amount of labelled data-sentiment analysis, developing relationships between text and large amounts of unlabelled data could be an example of this.

Step 5: Hyperparameter Optimisation

Hyperparameter optimisation is a crucial step in refining machine learning models, involving the systematic adjustment of parameters not learned during training. Techniques like grid search, random search and Bayesian optimisation are employed for the efficient exploration of hyperparameter combinations. Despite the computational cost, continuous evaluation and adjustment based on segmentation metrics are essential for identifying optimal values that lead to well-performing and generalisable models. In the context of medical image analysis this process is crucial for enhancing model accuracy.

Step 6: Model Evaluation

In evaluating machine learning models, employing multiple metrics is crucial for a comprehensive understanding of performance. Relying solely on a single metric can lead to skewed interpretations, as each metric captures different aspects of model behaviour. While a high F1 score may indicate robust performance, it might not account for nuances captured by other metrics like accuracy, which measures overall correctness, or specificity, which assesses the true negative rate.

In Medical Diagnostics Area Under the Curve (AUC) is a method which is useful in diagnostics as it performs well contexts where distinguishing between two binary states is crucial (e.g. diseased and healthy) and where the data may be imbalance. AUC is less useful in contexts where classification errors become more significant for instance incorrectly flagging a transaction as fraudulent (false positive) is much more significant in the finance industry than in medicine where over-diagnosing is not as much of an issue.

Thus employing cross-validation techniques is paramount for ensuring model reliability.

Step 7: Model Deployment

When considering deploying your ML model, it is vital that there is good cohesive teamwork between the computer scientists, domain specialists and data engineers. Continuous evaluation of your ML model is necessary, paying attention to Key Performance Indicators and metrics such as accuracy to evaluate model performance. Regular updates to the ML model, in line with new datasets and feedback, ensure that the ML model developed remain relevant and effective. Good documentation and auditing will be important for facilitating cross collaboration between different professionals. Auditing will also be important as calls for AI transparency, particularly important for those involved in public service, grow louder (Felzmann et al., 2020).

**Personal Perspective**

In this essay we have stressed the potency of ML to penetrate all parts of scientific and industrial life. Indeed we have gone a step further and highlighted how the phenomenon of ML is an unavoidable reality. We have also mentioned some of the ethical concerns with developing ML models. Unlike other industries, due to having a universal healthcare system, the UK government does have a large volume of control over healthcare data. It will be interesting to observe legislators and policy makers not only strike the balance between unleashing the power of AI and AI safety, but also manoeuvre in a way as to not be overtaken by adversarial actors who may be developing ML technologies but do not share the same ethical concerns (Montasari, 2023).

As the author looks ahead to developing ML Pipelines in the future, he notes the value in careful documentation. In the case that the product does not meet Key Performance Indicators he can work with or pass his work onto other professionals to work on. In the case of his product becoming successful he may find himself having to justify his decisions.

**Conclusion**

To draw this essay into concluding remarks, whilst ML is set to revolutionise modern life, for those who are immersing themselves in this field and wishing to successfully develop their own ML pipeline, they must be able to analyse their data, process the data and also choose a model which suits their end goals. This work may involve working with others from different domain expertise and different stakeholders, in such cases researchers should strive to make their work transparent for others to follow. Indeed researchers should also continue to monitor and fine tune their ML beyond deployment. In addition, to conclude, developers of ML Pipelines ought to consider how their work may impact the environments in which they are deployed.

**References**

Bank of England. (2022) Machine Learning in UK Financial Services. Available at: <https://www.bankofengland.co.uk/-/media/boe/files/report/2022/machine-learning-in-uk-financial-services.pdf> [Accessed: 15 February 2024].

Bilic, P., Christ, P., Li, H., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Humpire Mamani, G., Chartrand, G., Lohöfer, F., Holch, J.W., Sommer, W., Hofmann, F., Hostettler, A., Lev-Cohain, N., Drozdzal, M., Amitai, M.M., Vivanti, R., Sosna, J., Ezhov, I., Sekuboyina, A., Navarro, F., Kofler, F., Paetzold, J.C., Shit, S., Hu, X., Lipková, J., Rempfler, M., Piraud, M., Kirschke, J., Wiestler, B., Zhang, Z., Hülsemeyer, C., Beetz, M., Ettlinger, F., Antonelli, M., Bae, W., Bellver, M., Bi, L., Chen, H., Chlebus, G., Dam, E.B., Dou, Q., Fu, C.-W., Georgescu, B., Giró-i-Nieto, X., Gruen, F., Han, X., Heng, P.-A., Hesser, J., Moltz, J.H., Igel, C., Isensee, F., Jäger, P., Jia, F., Kaluva, K.C., Khened, M., Kim, I., Kim, J.-H., Kim, S., Kohl, S., Konopczynski, T., Kori, A., Krishnamurthi, G., Li, F., Li, H., Li, J., Li, X., Lowengrub, J., Ma, J., Maier-Hein, K., Maninis, K.-K., Meine, H., Merhof, D., Pai, A., Perslev, M., Petersen, J., Pont-Tuset, J., Qi, J., Qi, X., Rippel, O., Roth, K., Sarasua, I., Schenk, A., Shen, Z., Torres, J., Wachinger, C., Wang, C., Weninger, L., Wu, J., Xu, D., Yang, X., Yu, S.C.-H., Yuan, Y., Yue, M., Zhang, L., Cardoso, J., Bakas, S., Braren, R., Heinemann, V., Pal, C., Tang, A., Kadoury, S., Soler, L., van Ginneken, B., Greenspan, H., Joskowicz, L., Menze, B., (2023) 'Prospects and Challenges of the Machine Learning and Data-Driven Methods for the Predictive Analysis of Power Systems: A Review', Energies, 16, 4025. <https://doi.org/10.3390/en16104025>

David, A.N., Sewsynker-Sukai, Y., Meyer, E.L., and Gueguim Kana, E.B. (2023) 'Harnessing Artificial Neural Networks and large language models for bioprocess optimization: Predicting sugar output from Kraft waste-based lignocellulosic pretreatments', Industrial Crops & Products, 206, 117686. Available at: <https://doi.org/10.1016/j.indcrop.2023.117686> [Accessed: 15 February 2024].

Gartner. (2023) 'What's New in Artificial Intelligence from the 2023 Gartner Hype Cycle'. Gartner. Available at: <https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2023-gartner-hype-cycle> [Accessed: 15 February 2024].

Goodell, J.W., Kumar, S., Lim, W.M., and Pattnaik, D. (2021) 'Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis', Journal of Behavioral and Experimental Finance, 32, 100577. Available at: <https://doi.org/10.1016/j.jbef.2021.100577> [Accessed: 15 February 2024].

Grand View Research. (2023). AI In Medical Imaging Market Size, Share & Trends Analysis Report By Technology (Deep Learning, NLP), By Application (Neurology, Respiratory & Pulmonary, Cardiology), By Modality, By End-use, By Region, And Segment Forecasts, 2023 - 2030. Available at: <https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-medical-imaging-market> [Accessed 27 Jan 2024]

Felzmann, H., Fosch-Villaronga, E., Lutz, C., Tamò-Larrieux, A. (2020) 'Towards Transparency by Design for Artificial Intelligence', Science and Engineering Ethics, 26, pp. 3333–3361. <https://doi.org/10.1007/s11948-020-00276-4>

Khan, U. (2024) 'Machine Learning Pipeline', Queen Mary University of London, [image]

ML Cheatsheet. 2017. Glossary. Available at: <https://ml-cheatsheet.readthedocs.io/en/latest/glossary.html> [Accessed: 15 February 2024].

Montasari, R. (2023) 'National Artificial Intelligence Strategies: A Comparison of the UK, EU and US Approaches with those Adopted by State Adversaries', in Countering Cyberterrorism. Advances in Information Security, vol 101. Cham: Springer. Available at: <https://doi.org/10.1007/978-3-031-21920-7_7> [Accessed: 15 February 2024].

Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K., & Tekade, R.K. (2021) 'Artificial intelligence in drug discovery and development', Drug Discovery Today, 26(1), pp. 80-93. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7577280/> [Accessed: 15 February 2024].

Strielkowski, W., Vlasov, A., Selivanov, K., Muraviev, K., & Shakhnov, V. (2023) 'Prospects and Challenges of the Machine Learning and Data-Driven Methods for the Predictive Analysis of Power Systems: A Review', Energies, 16(4025). <https://doi.org/10.3390/en16104025>

Vajjala, S., Majumder, B., Gupta, A. and Surana, H. (2020) Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems. O'Reilly Media.

Velarde, G. (2019) 'Artificial Intelligence and Its Impact on the Fourth Industrial Revolution: A Review'. International Journal of Artificial Intelligence & Applications (IJAIA), 10(6), November 2019. Available at: <https://arxiv.org/ftp/arxiv/papers/2011/2011.03044.pdf> [Accessed: 15 February 2024].