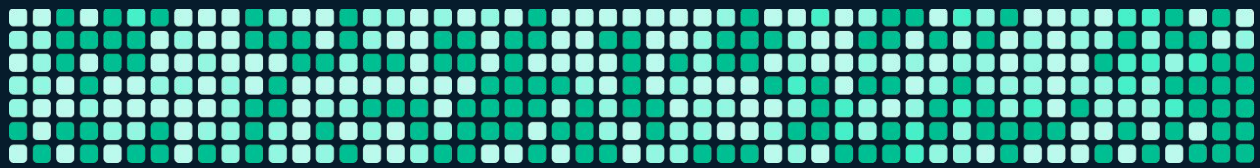
loxz digital

Q3 2022

ML Readiness Survey Report



By Miu Lun (Andy) Lau
Data Scientist, Loxz Digital

loxz digital

Preface

For the past five years, we have seen tremendous growth and changes to the Machine Learning / Artificial Intelligence's landscape. We have seen development and deployment of many different ML systems and frameworks. At Loxz Digital, we observed many different adaptations of ML related techniques in a wide variety of fields such as business/finance, health science, and many others including retail. We have also observed the proliferation of Data and AI machine learning tools which enables business to accelerate their deployment and training for ML applications. As such, the following report will be an overview of the popular tools used in ML as well as their individual use case.

Frameworks

Machine learning frameworks are typically used as a rapid basis for model and training development. Over the years, there have been multiple ML frameworks that garner significant users' traction owing to their usability, performance, and ease of development. According to a survey performed by Kaggle in 2021 over the ML framework popularity: Scikit-learn [8] still dominates as a majority in ML since it provides numerous simple to use functions for pre-processing. In second place we have Tensorflow being the most dominant. Tensorflow [6] is a very well documented framework with many tutorials and pre-trained models available. It also provides a plethora of analysis and visualizations tools for users to explain their ML findings. Explainable AI has become an industry of itself but the current investments in Generative AI is becoming more commonplace and will evolve into a much more popular industry as we see it. Furthermore, the high level API wrapper (Keras [2]) also provides users the ability to rapidly develop new models on the fly with low level abstractions.

loxz digital

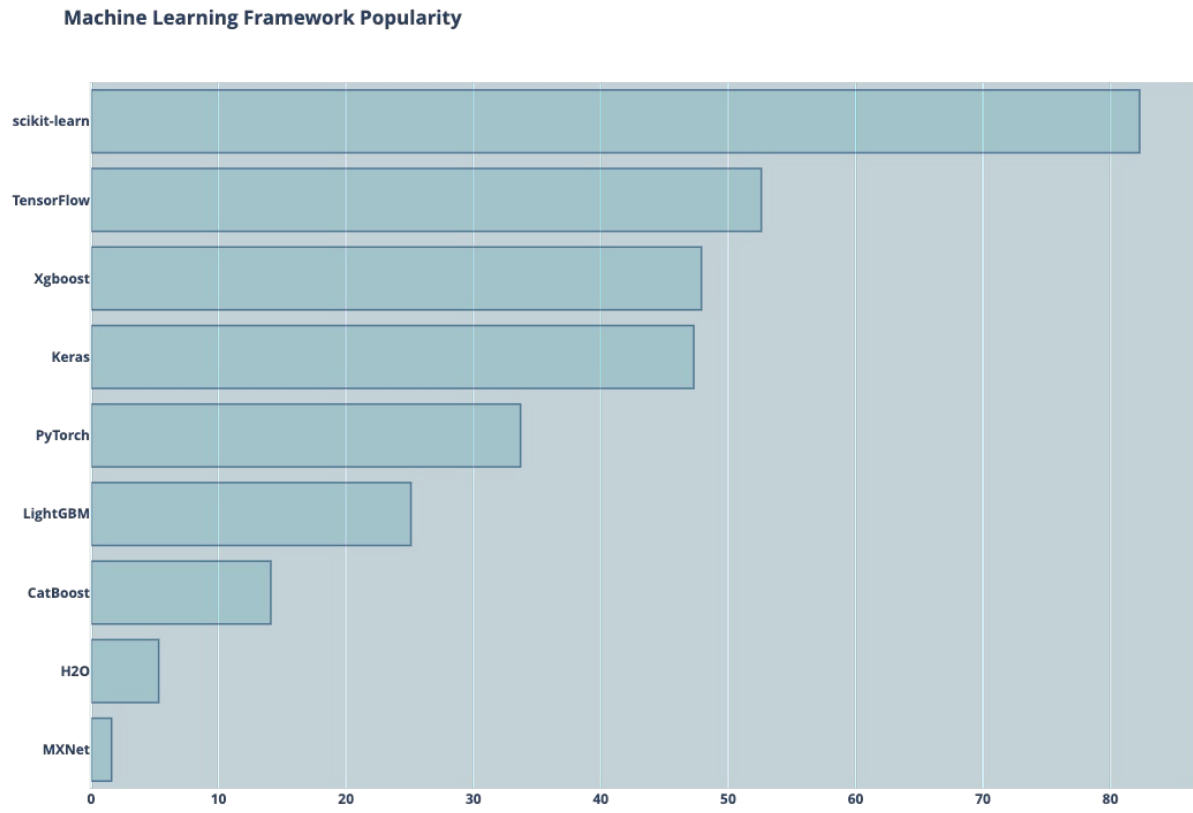


Figure 1. ML framework popularity (https://plotly.com/~laumiulun/60/)

On the other hand, PyTorch [7] developed by Facebook offers an alternative to TensorFlow. Pytorch is built to mirror the feel of python programming language and offers advantages such as Data-Parallelism which can utilize multi GPU effectively as a batch process. From there, you can observe other boosting frameworks included in this list from Xgboost [1], CatBoost [3] and LightGBM. The goal of boosting is creating an accurate classifier from a weak classifier by subdividing the training data and using each part to train different models or one model with a different setting, and then the results are accumulated together using a majority vote.

At Loxz Digital, our data scientists and engineers utilize a combination of Tensorflow, Keras and PyTorch to create our framework. We also utilize many features from each of the framework to create and construct our accurate models. Most of the models are built in Jupyter Notebooks. Sagemaker is also the tool of choice for managing the models and as well as versioning. We also will use Sagemaker for deployments. Most of the models in our portfolio are now in version 2 as we continue to add features to the models for higher accuracy and usability as we receive feedback from the field.

Data Management

The core of machine learning relies heavily on the source of data, as well as the validity and quality of the data. Most of the time, more data is not the answer, but clean labeled data alongside the right data is more important. However, data creation, labeling, and cataloging still relies on a combination of traditional techniques and **man-hours** which significantly reduces throughput. It is still costly to produce high quality models. Recently there have been many new packages and frameworks developed related to data management. Synthetic data is used to perform and evaluate model performance without compromising safety and efficiency. Many new companies such as Hazy, Mostly-AI, and YData aim to create artificial synthetic data to perform data analysis on machine learning related problems.

At Loxz Digital, we have also generated synthetic data to be used for training our models related for our email campaign predictions. We have applied synthetic data along with data augmentation techniques to our models to generate real time predictions for email related campaigns. Our model can be optimized for Call to Actions, Character Counts, Sentiment Analysis and Image Optimization, and many more email metric characteristics. We are also in the process of partnering with large email service providers to utilize their data to train, predict, and verify all our existing models. A future vision of one of our models, “sentiment analysis” will have an additional input regarding the “intensity” of the sentiment. We are working on this feature currently. The intensity range might be from 1-5 and produces language that intensifies the sentiment. We hope to use Generative AI to produce proper language for the campaign engineer prior to deployment.

Data Labeling

Data labeling is also a fundamental throughput problem in ML, where it relies on human related factors to generate a tagged dataset. As a result, there has been software developed for labeling jobs. The most popular and open sourced is ***Label Studio*** **[9]**, where users can define and provide their data and *Label Studio* to segment into correct bins for users to output labeled data. Additionally, there are other pay services such as **Amazon SageMaker Ground truth,** which couples with Amazon mechanical Turks to perform human labeling tasks.

At Loxz Digital, we have used a combination of personal data labeling from confidential data from email service providers as well as Amazon SageMaker ground truth to create and generate large, high-quality, labeled dataset. We all know that trained models using proprietary datasets differ from models that are productionalized. Much as to go right to maintain the accuracy of a model transitioning from testing to production. While many models are retrained sparingly, (once a month, or once a year) we plan to retrain our models weekly with fresh data. There will come a time when retraining our models in real-time, regularly. Data drift and algorithmic drift are commonplace in ML. In order for our models not to breach thresholds, retraining will paramount for all the models in production.

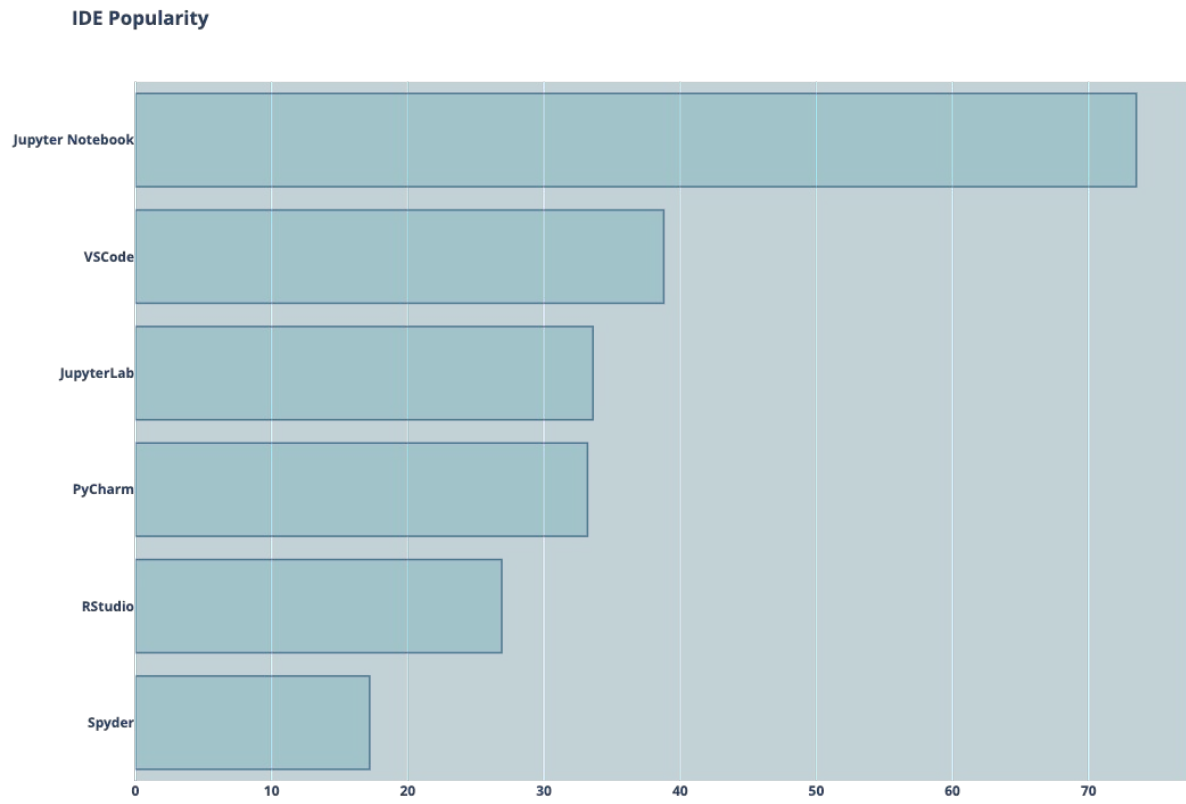


Figure 2. IDE popularity (https://plotly.com/~laumiulun/62/)

Notebook and IDE

Notebook and Integrated Development Environment (IDE) are a paramount part of a data scientist's toolkits. It allows for users to observe changes in the model performance and allows for rapid modifications to adjust the quality. From the same survey from Kaggle, it is observed that jupyter notebook and VScode exists as the most popular IDE in 2022.

At Loxz Digital, we utilize Amazon Sagemaker as the main IDE since it permits active collaboration between our remote and talented data scientists. Jupyter notebooks and endpoints are also established in our AWS envirnoment.

Models Algorithms

While over the past decades, there have been many model architectures that came and went, and their popularity stems from their ability to perform certain analyses or predictions. Related to machine learning, the simplest algorithms can be regression related algorithms such as linear or logistic regression. These algorithms are applied daily across a wide variety of different domains.

Decision Tree algorithms are a type of popular algorithm for classification tasks. It aims to decide the sample population into two or more homogeneous sets based on the most significant attributes and independent variables. Support Vector Machines (SVM) are also very popular which aims to classify problems using dimensional vectors.

With regards to Natural Language Processing (NLP) and language related models, it requires the processing of sequential or sets of data (sentences) which relate various sentence structures to be applied. Recurrent Neural Network (RNN) were mainly applied before because it can create weights of significant parts of input for later recognition, also known as Long Short Term Memory (LSTM). RNN lacks the ability to parallelization and coupling with other models since it relies on the data from the previous node, therefore the serialization is a detriment.

Generative AI and Transformers

Currently in the machine learning landscape, the largest development comes mainly from the progress of generative models which are used to provide AI generated Art. The first model to generate text-to-image comes from OpenAI DALL-E [12] which uses a Generative Pre-Trained transformer (GPT) model. However since DALL-E was closed source, many alternatives have been proposed instead such as **Stable Diffusion** [13] or MidJourney [14]. The text-to-image model relies on a large dataset consisting of billions of image and captions pairs. Additionally, it uses a diffusion related model called latent diffusion which aims to remove gaussian noise on training images.

Investors are devouring Generative AI companies. Several early stage companies are raising tens of millions since we believe this is the next frontier for AI/ML organizations. It is in Loxz' best interest to collaborate with a few of these companies, and to offer our ecosystem predictive analytics on text to image creatives. This is our goal and will impact our image optimization models including our data augmentation model. It stands to reason that when a campaign engineer creates her image in the email editor, uploads that image and runs a model, we should be able to provide predictive analytics on "text to image" model she just created. These are still very early days in generative AI, but certainly important to say the least.

Data Visualizations & Debugging

Data visualizations are fundamental and an integral role in machine learning. It is used to explain and explore the models' output. The five visualization factors in Machine Learning includes: *Explainability*, *Debugging & Improvements*, *Comparisons & Selections*, *Articulate Concepts*. Using a variety of popular tools in the python libraries such as ggplot2 [11], Plotly [5], Matplotlib [4], Seaborn [10], and etc., we can create visulaizations to explain model architecture, parameters performance, model metrics, and many other parameters which helps provide metrics for how well our model is performing under different conditions. It also gives data scientists a way to track the performance of the model as new data is being ingested.

At Loxz Digital, we utilize matplotlib for our internal model debugging and optimization, and our front-end **HuggingFace** utilizes Plotly for interactive visualizations. Additionally we also utilize other popular libraries such as Bokeh for some of the model and metrics visualizations. Visualization is also used in Grafana for model monitoring. More on Grafana in a later report, due out in late January 2023.

loxz digital

About Us

Loxz Digital Group is a Machine Learning Collective located in Berkeley, CA. Established in December of 2020, Loxz is focused on serving RealTime predictive analytics. We supply models and serve predictions within smart workflows to clients of the Email Service Provider network and enterprises and are in discussions with serving location specific predictions to law enforcement to reduce gunshot violence. We employ a servant-leadership management style where every employee or advisor has a distinct voice. Specifically, RealtimeML is at the bedrock of what we do. Collectively, the current assembled team has over 40 years of ML experience, housing 8 data scientists, all located in the United States and Canada.

Visit www.loxz.com

References

[1] Tianqi Chen and Carlos Guestrin. “XG Boost: A Scalable Tree Boosting Sys tem”. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD ’16. San Francisco, California, USA: ACM, 2016, pp. 785–794. ISBN: 978-1- 4503-4232-2. DOI: 10.1145/2939672. 2939785. URL: http://doi.acm.org/10. 1145/2939672.2939785.

[2] Francois Chollet et al. Keras. 2015. URL: https://github.com/fchollet/keras.

[3] Anna Veronika Dorogush, Vasily Er shov, and Andrey Gulin. “CatBoost: gradient boosting with categorical features support”. In: arXiv preprint arXiv:1810.11363 (2018).

[4] John D Hunter. “Matplotlib: A 2D graph ics environment”. In: Computing in sci ence & engineering 9.03 (2007), pp. 90– 95.

[5] Plotly Technologies Inc. Collaborative data science. 2015. URL: https://plot.ly. [6] Martín Abadi et al. TensorFlow: Large Scale Machine Learning on Heterogeneous Systems. Software available from ten sorflow.org. 2015. URL: https://www. tensorflow.org/.

[7] Adam Paszke et al. “PyTorch: An Im perative Style, High-Performance Deep Learning Library”. In: Advances in Neu ral Information Processing Systems 32. Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/ paper / 9015 - pytorch - an - imperative - style-high-performance-deep-learning library.pdf.

[

loxz digital

8] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: Journal of Machine Learning Research 12 (2011), pp. 2825–2830.

[9] Maxim Tkachenko et al. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label studio. 2020-2022. URL: https://github . com/heartexlabs/label-studio.

[10] Michael L. Waskom. “seaborn: statistical data visualization”. In: Journal of Open Source Software 6.60 (2021), p. 3021. DOI: 10.21105/joss.03021. URL: https: //doi.org/10.21105/joss.03021.

[11] Hadley Wickham. “Data analysis”. In: ggplot2. Springer, 2016, pp. 189–201.

[12] DALLE. “https://openai.com/blog/dall-e/”

[13] Robin Rombach et al. “High-Resolution Image Synthesis With Latent Diffusion Models”. In: Proceedings of the IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 2022, pp. 10684–10695.

loxz digital

Contributors

Miu Lun (Andy) Lau, Data Scientist,
Lead Author, Lead Analyst

Yumi Koyanagi, Designer
Report Designer