A Data-Centric Approach to Improve Machine Learning Model's Performance in Production

Pritom Bhowmik, Arabinda Saha Partha

***Abstract:** Machine learning teaches computers to think in a similar way to how humans do. An ML models work by exploring data and identifying patterns with minimal human intervention. A supervised ML model learns by mapping an input to an output based on labeled examples of input-output (X, y) pairs. Moreover, an unsupervised ML model works by discovering patterns and information that was previously undetected from unlabelled data. As an ML project is an extensively iterative process, there is always a need to change the ML code/model and datasets. However, when an ML model achieves 70-75% of accuracy, then the code or algorithm most probably works fine. Nevertheless, in many cases, e.g., medical or spam detection models, 75% accuracy is too low to deploy in production. A medical model used in susceptible tasks such as detecting certain diseases must have an accuracy label of 98-99%. Furthermore, that is a big challenge to achieve. In that scenario, we may have a good working model, so a model-centric approach may not help much achieve the desired accuracy threshold. However, improving the dataset will improve the overall performance of the model. Improving the dataset does not always require bringing more and more data into the dataset. Improving the quality of the data by establishing a reasonable baseline level of performance, labeler consistency, error analysis, and performance auditing will thoroughly improve the model's accuracy. This review paper focuses on the data-centric approach to improve the performance of a production machine learning model.*

***Keywords:** Annotation, Augmentation, big-data, bias-error, baseline, consistent-leveling, Data-centric, model-centric, error-analysis, good-data, Model-accuracy, Human-level-performance, proxy.*

I. INTRODUCTION

In academic and research settings, traditional ML modelling is less complicated. Typically, some standard datasets are supplied, and they are most often cleaned and labeled. So, that gives an ML model, which makes good predictions.

However, developing an ML model that is production-ready and performing post-deployment improvements are extensively iterative processes. In the machine learning life cycle, investing most resources in improving the data, model access gives the highest return. Big data makes it possible to feed ML models an enormous amount of data, which improves ML models' performance.

Nevertheless, after achieving a certain level of accuracy, the big data does not help much to improve the performance further. That gives the reason is to shift from the "Big data" to "Good Data" concept. An enormous amount of data, aka big data, helps the model avoid overfitting, but that cannot achieve 98-99% accuracy. Furthermore, most state-of-the-art ml models today require a higher level of accuracy to make them ready to deploy in production. Moreover, safety-critical ml models like autonomous cars, co-pilot systems, medical models, and robotic technology demand around 99.9% accuracy to make them suitable for production. Furthermore, we need a consistent and correctly labeled dataset, aka good data with a state-of-the-art model, to achieve this purpose. While data is collected from various sources using a data pipeline, it needs to go through extensive data cleaning and formatting processes. Moreover, these processes can filter and clean the data for a certain level, making them suitable for feeding a machine learning model. We can use those data to build a model and almost get around 60-70% accuracy. However, for developing a 95-99% accuracy model, we must need a correctly organized, formatted, sorted, and labeled dataset.

Furthermore, for the unstructured datasets, this process is much more challenging. Applying correct data formats, labeling consistently, and establishing a baseline and human-level performance are very iterative processes and frequently need to be changed during pre-production and post-production developments. Developing the code (algorithms) and tuning hyperparameters, regularization, and optimization processes are also very iterative processes, but after a certain point of model development, we get the desired machine learning model with 75-80% accuracy. That is a good accuracy score, but there is a demand to go beyond the 90% accuracy level for most problem use-cases. And without an excellent quality of data, this will be impossible to achieve. Data scientists & ML engineers often face this problem where they have to switch from big data concepts to good data concepts. Data collection, processing, and labeling tasks are 70% of the actual work of an ML project. Moreover, to develop a model production-ready, the data-centric approach is very crucial.

Manuscript received on October 09, 2021.
Revised Manuscript received on October 27, 2021.
Manuscript published on October 30, 2021.
* Correspondence Author

Pritom Bhowmik*, B.Tech. Department of Computer Science & Engineering, Institute of Engineering & Management, Salt-Lake, Kolkata, India. Email: pritom01bh@gmail.com

Arabinda Saha Partha, B.Tech. Department of Computer Science & Engineering, Institute of Engineering & Management, Salt-Lake, Kolkata, India.Email: official.neo.partha@gmail.com

II. LABELER CONSISTENCY

One of the critical tasks in the data preparation process is consistent labeling. For unstructured data like text, audio, image, and video, it is important to properly categorize and annotate data for a specific use case. In many use-cases, there may need to label more than 10000 unstructured data. And a dedicated labeler team may work on that particular task. In a specific scenario, maintaining labeling consistency is very crucial. The example below shows inconsistent labeling where several labeler teams label the same kind of data differently. And this will reduce the data quality intensively. The consistency of the data is paramount.

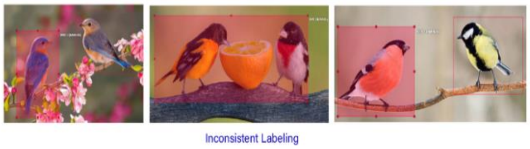


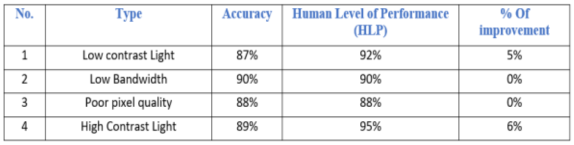
Figure 1: Example of inconsistence Labeling

For audio data, there may have traffic, plane, human noise in the background. Avoiding background noise is not a good way to label audio data. And we must follow a fixed baseline to label those data. In a complicated dataset, the labeling task becomes most crucial for achieving the desired accuracy. Having inconsistent labels in the dataset causes some critical errors. An MIT study found systematic labeling errors in the most popular AI benchmark datasets like CIFAR with around 5.83% incorrect labeling. Amazon review dataset also contains 4% label errors. These incorrect labeling causes errors like mislabelled images, mislabelled test sentiments, like a positive product review described as a negative review, and mislabelled audio of YouTube videos. Furthermore, the consequences may cause some censorious and unethical results.

III. ESTABLISHING BASELINE

Establishing a baseline level of performance is always a crucial step to improve the quality of data. Most ML models work on massive datasets, and improving all the data is not straightforward. It requires time, expertise and may be very costly sometimes. Establishing a baseline provides the possibility of improvements in the datasets. For an unstructured dataset, the human level of performance is the best way to establish a baseline. The table below provides the percentage of improvement that is possible in a specific ML model. This table is for a visual inspection model that detects an unauthorized person entering an area.

Table 1: Baseline Table



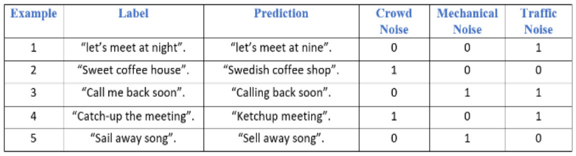
No.	Type	Accuracy	Human Level of Performance (HLP)	% Of improvement
1	Low contrast Light	87%	92%	5%
2	Low Bandwidth	90%	90%	0%
3	Poor pixel quality	88%	88%	0%
4	High Contrast Light	89%	95%	6%

The dataset used for the visual inspection model has different types of labeled images like low contrast light images, low bandwidth, poor pixel quality, and high contrast light images. When the model reaches a certain point where we have achieved an accuracy, around 70-75%, we should go for a data-centric approach where we will focus on improving a particular type of data with a good percentage of improvement possibility compared to human-level performance (HLP). In the above example, low contrast and high contrast light image data have huge room for quality improvement. The idea is not about exceeding the HLP but building a model that can perform as well as humans. Low bandwidth and poor pixel image data have no room for improvement because those types of accuracies are equal to the HLP. So, working with data, that has a good percentage of improvement, improving data quality will improve the overall accuracy of the machine learning model. This process will help establish a baseline. And the baseline will help to indicate what might be possible.

IV. ERROR ANALYSIS

Every step of a machine learning project is an iterative process. And by analyzing errors, we can identify which algorithm works most effectively. But in a data-centric approach, we do error analysis to improve the data iteratively. Error analysis process identifies which types of data do poorly in the algorithm. The following table gives an overview of different types of data's performance in a speech recognition model. By performing the error analysis, we know that we should prioritize traffic noise data where most of the wrong prediction occurs.

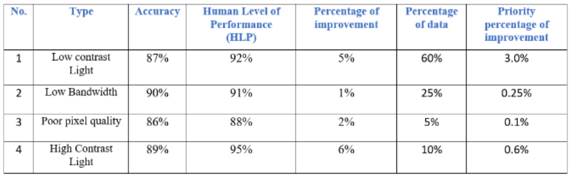
Table 2: Error Analysis Table 1



Example	Label	Prediction	Crowd Noise	Mechanical Noise	Traffic Noise
1	"let's meet at night".	"let's meet at nine".	0	0	1
2	"Sweet coffee house".	"Swedish coffee shop".	1	0	0
3	"Call me back soon".	"Calling back soon".	0	1	1
4	"Catch-up the meeting".	"Ketchup meeting".	1	0	1
5	"Sail away song".	"Sell away song".	0	1	0

While working with a large amount of data, prioritizing what data to work on is always a good practice. The main objective is to decide on the most important categories to work on. Sometimes there may have a big gap between the model's accuracy and human-level performance (HLP). However, how much that type of data is present in the dataset is a critical consideration.

Table 3: Error Analysis Table 2



No.	Type	Accuracy	Human Level of Performance (HLP)	Percentage of improvement	Percentage of data	Priority percentage of improvement
1	Low contrast Light	87%	92%	5%	60%	3.0%
2	Low Bandwidth	90%	91%	1%	25%	0.25%
3	Poor pixel quality	86%	88%	2%	5%	0.1%
4	High Contrast Light	89%	95%	6%	10%	0.6%

The above table shows the priority percentage of improvement, which identifies which type of data has the most room for improvement.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Low contrast light type images are 60% of the total dataset, and there is a 5% gap between accuracy and HLP. So, improving that category will improve the overall performance of the model. The ML team must several questions during that particular task, such as how much room for improvement there is, how frequently that category appears and how easy it is to improve accuracy in that category. After prioritizing a specific category, we can collect more data or generate more data using the data augmentation method.

V. DATA AUGMENTATION

If we have many parameters, we would like to give the machine learning model a proportional number of examples to perform better. The data augmentation technique is used to extract more information or data from the original dataset through augmentations. These augmentations artificially inflate the training dataset size by using data oversampling or data wrapping.

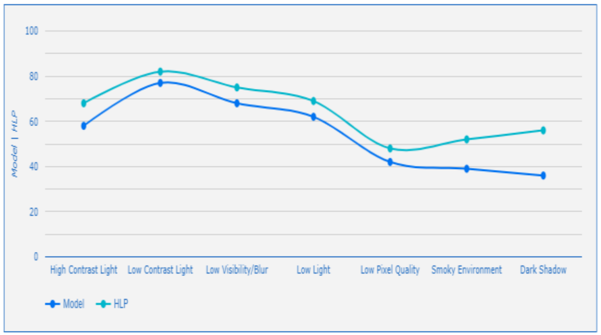


Figure 2: Data Augmentation Model vs HLP

Oversampling augmentations create synthetic instances and add them to the training dataset. The graph represents the possible improvement in a different scenario that can overall improve the performance model. In many cases, there are a limited amount of data collected, and after splitting data for validation and test, there remains significantly less data for training the model. During the training of a machine learning model, we need to tune its parameters so that it can map an input (say, audio) to some output(label). And we need a lot of parameters so that we can feed the machine learning model a proportional number of examples to achieve a higher level of accuracy. In the graph, we can get the overview that we need more examples in smoky and dark environment conditions because there are big gaps between Human-level performance (HLP) and the model.

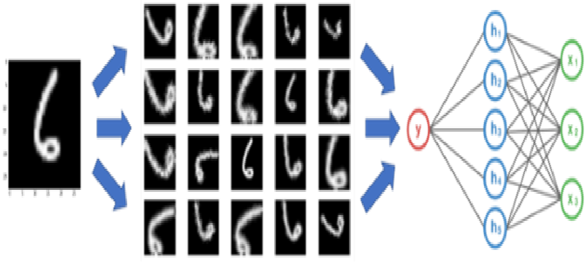


Figure 3: Neural Network

In a real-world scenario, we may have a dataset of images in a minimal set of conditions. But the output can be in various conditions, like different brightness, time, weather, scale, or location. So, we add synthetically modified data (Images) to the machine learning model to account for these situations and improve the overall performance.

VI. HUMAN-LEVEL ERROR AS A PROXY FOR BAYES ERROR

The human-level error gives an estimation for the Bayes error with the help of error analysis. For detecting a disease, we can fix the human-level error with the help of typical doctors rather than a very experienced team of doctors. But to serve human-level error as a proxy for Bayes error, we must go for the best possible way to fix the baseline. A team of regular doctors may have 1% error and an experienced doctor's team may have 0.4% of error. So, the base error should be < 0.4% or equal. But depending on the use-case of the problem, we can adjust the baseline for the Bayes error. It is not so important which one we choose for Bayes error. Because the gap between training error and Bayes error defines as avoidable Bayes, needs to be reduced by using a bias reduction technique to improve the performance of the model. But if the gap between Bayes error and training error is less than 1%, but the gap between Dev error and training error is more than 1%, we need to focus on reducing variance error. Again, if the training error is around 1 %, we should fix the bias error as low as possible to improve performance. So, establishing the human-level performance threshold is a big step to improve the performance of a machine learning model.

VII. PERFORMANCE AUDITING

Performance auditing is the very last line of defense before deploying the model in production, and this can help us to avoid post-deployment errors. A model can do very well in accuracy, F1 scores, or other error analysis processes, but performance auditing helps to check for fairness/bias and its performance in rare situations. The model can perform poorly or unethically on a particular subset of data and produce certain common errors. A facial recognition system can perform very well but, in the dataset, we may miss adding certain races of human faces (e.g., African).So, the system can have around 99% accuracy and perform very well in production. However, it can cause some issues while detecting black faces. And that raises an ethical problem that should be avoided before deploying in production. Performance auditing helps engineers to avoid such problems by detecting them before deploying the model in production. The most common way to practice performance auditing is establishing metrics to assess performance against these issues on the appropriate subset of data (e.g., ethnicity, gender).

VIII. CONCLUSION

The data-centric approach is not practiced extensively in academia and research, but it is crucial for developing ML products and making them production-ready. In any AI system, data and models both work hand in hand to produce the desired result. AI research has been model-centric in nature. Big data helps the Machine learning model to perform immensely well and to avoid overfitting. But with the demand for improving AI products, more accurate than a human, it becomes challenging to improve performance only by changing the model(code).

"Examining a sample of recent publications revealed that 99% of the papers were model-centric with only 1% being data-centric. - Andrew Ng"

And the systematic change in the dataset is improving the performance of a machine learning model substantially. Hopefully, future AI research will be directed to data-centric along with model-centric.

REFERENCES:

1. https://venturebeat.com/2021/03/28/mit-study-finds-systematic-labeling-errors-in-popular-ai-benchmark-datasets/
2. https://www.youtube.com/watch?v=06-AZXmwHjo
3. https://www.deeplearning.ai/
4. https://www.coursera.org/learn/introduction-to-machine-learning-in-production
5. https://bardhrushiti.medium.com/human-level-performance-and-bayesian-optimal-error-fadf4f55cd48
6. Machine Learning Algorithms to Improve Model Accuracy and Latency, and Human-Autonomy Teaming by Vincent Houston, Bryan Barrows, Walter j Manuel, Lisa le Vie.
7. Data Acquisition for Improving Machine Learning Models by Yifan Li, Xiaohui Yu, Nick Koudas
8. Hands-on-Machine-Learning by Aurelien Geron
9. Towards better analysis of machine learning models: A visual analytics perspective by Shixia Liu, Xiting Wang, Mengchen Liu, JunZhu
10. Advanced machine learning model for better prediction accuracy of soil temperature at different depths
11. https://www.dummies.com/programming/big-data/data-science/10-ways-improve-machine-learning-models/
12. https://www.dummies.com/programming/big-data/data-science/performing-classification-tasks-machine-learning/
13. https://en.wikipedia.org/wiki/Machine_learning
14. https://en.wikipedia.org/wiki/Training,_validation,_and_test_sets#training_set
15. https://en.wikipedia.org/wiki/Regularization_(mathematics)
16. https://en.wikipedia.org/wiki/Loss_functions_for_classification
17. https://en.wikipedia.org/wiki/Data_augmentation
18. https://en.wikipedia.org/wiki/Oversampling_and_undersampling_in_data_analysis#Oversampling_techniques_for_classification_problems
19. https://en.wikipedia.org/wiki/Sampling_(statistics)
20. https://en.wikipedia.org/wiki/Probability_distribution

AUTHORS PROFILE

Pritom Bhowmik, B.Tech. (Computer Science & Engineering) Institute of Engineering & Management, Salt-Lake, Kolkata, India Research work: Research focused on data science for business and big-data. Recent works are on data-centric AI approach and analysis of uncertainty and detecting anomalies in high-dimensional big data. Previous research paper was published on artificial neural network. Email**:** pritom01bh@gmail.com

Arabinda Saha Partha, B.Tech. (Computer Science & Engineering) Institute of Engineering & Management, Salt-Lake, Kolkata, India Research Work: Computer Networking and Cyber Security. Researching on Distributed database management system & AI for Cy Security. Email: official.neo.partha@gmail.com