Data Validation Process in Machine Learning Pipeline

Ram Mohan Vadavalasa
India

Abstract— The Machine learning is a powerful tool for finding patterns from massive amounts of data. Data is being generated, collected, transformed, processed, and analyzed for machine learning end to end life cycle. Machine learning research has focused on improving the accuracy and efficiency of training algorithms, but there is an equally important problem of monitoring the quality of data fed into machine learning. Machine learning pipeline treats training and serving data as an important product asset, which is equal to the algorithm and infrastructure used for learning. Validating data is an essential requirement to certify the worthiness and benchmark of the Machine Learning system.

Keywords: Machine Learning Pipeline, Data Validation Process

Purpose of the paper:

Data is collected from different resources with a fast velocity, so that in initial phase quality of data is imperfect. Data with poor quality is determined by the probability of impact of inefficient data on the result of the machine learning model. Detecting errors in the early stages of pipeline is better for model data quality and possible to save enough resources from debugging issues later in the pipeline. In this paper, we present a data validation system that is able to detect and notify, if there is any inconsistency and abnormality specifically in the data fed into the machine learning pipeline

I. INTRODUCTION

Machine Learning (ML) is widely used to investigate insights from massive amounts of data.

In industry, retail, health care, banking, genomics, government, education, and self-driving cars use of data analytics and machine perception is increasing significantly. Almost every large scale company is increasing investments in data analytics and would like to utilize data to make better business decisions.

Collecting large amounts of data [14] from different sources like system logs, browsing data, sensor data and later storing this raw data in big storage platforms for further machine learning processings.

For example, web application management is using visited users in log data for analyzing the user behavior, on which page users spend a long time, on which topic users showing more interest; using this raw data from customers able to predict future usage and customer preferences.

In companies data is generated from heterogeneous sources and the quality of data is far from perfect and this poor data quality will affect the quality of the generated machine learning model and it finally affects company decisions.

Error-free data will help to understand the patterns [18,21] in the data and effects the final output of the complete machine learning model. Additionally, based on generated predictions from the machine learning model will become a source for generating more new data for the next machine learning cycle. Keeping this cycle in mind, initially taking enough care for producing error-free data; because eventually even small error in the data could gradually effects model performance over a period of time.

Therefore, it is imprerative to catch errors in the data in early stage because before they propagate through the machine learning pipeline and cause a biased model.
Considering above all observations we need to hoist data in ML pipeline for building infrastructure with continuous monitor and validate data throughout different stages and cycles of ML pipeline.

Data validation is not a unique problem in ML but it contains lot of challenges to meet our requirement for model generating. Filtering errors earlier in the data is an important task, as it helps to debug the root cause and possible to analyze data in the initial phase and it is useful for getting a complete overview on the data.

The Data validation is an important step to improve data quality and filter un-related data from data storage facilities.

This paper focus on the problems of validating the input data fed into ML pipelines and to overcome errors and vulnerabilities in the data.

II. RELATED WORK:

Understanding and preparation data initially is an important parameter in the ML life cycle.

Collecting data from diversified sources and integrated together to form a single source for data processing. But this collected data from mixed sources contains a lot of anomalies, which are later in the pipeline causes irregularity in the model. Data is required to pre-process for increasing data quality for model generation.

Input is transformed into the training model after the raw data is prepared for the processing. Raw input data is converted to a specific format of features and values, which can be further useful for model training.

For example, the data is mapped to a vector product of three values. The key question to ask for data preparation are, what features we can generate from data, what are the outcomes of the feature values, what are the best implementations to transcode those values.

During the initial development of a model, data preparation flows down to the model pipeline, once the model is matured, the model focal point shift towards latency reduction and resource optimization, by just simply prioritizing main features in the model and keeping the same accuracy.

Improving data quality and accuracy of the generated model is important because once the model data delivered to production, new data will flow and join with existing data with different transformations for the next ML pipeline cycle. This data should be normalized to a standard representation for analyzing, what to expect from the data.

III. IMPLEMENTATION PROCESS

The main aim of data validation continuously checks and monitor the data quality in the ML life cycle. In ML life cycle [3] complete anomalies and errors should be detected before data processing into model training.

Finding errors in the data is a challenging task because Terabytes of data contains thousands of features and this data has the possibility to have a lot of errors.

Insufficient data in the ML is due to the unavailability of enough tools sets for data validation and data cleaning compared to traditional software processing. These errors in the data cause risks in the model and risks lead to negative consequences in the long run.

A lot of statistical testing and analyzing models are available for data analysis, some of them are aggregations, t-test, homogeneity tests, Z-test A/B testing, Chi-Square Test, correlation analysis, Wilcoxon Test, and time series analysis.

Data and their values are checked respective to their defined schemes. For example, hypothesis testing is explained here in order to present the relation between data and its respective defined schemes.

For example, we want to find statistics about a group of companies and their behavior in a specific field but in the real world it's hard to find complete statistics and variables about these companies, at this point we use statistical measures using hypothesis testing in order to estimate [8] their behavior. For example, their sample Mean ($\bar{x}$), variance (S^2), Standard deviation (S), Proportion (p), sample distribution, and standard error.

There are different hypothesis testing techniques are available such as, Simple Hypothesis, Complex Hypothesis, Empirical Hypothesis, Null Hypothesis ("H0"), Alternative Hypothesis ("H1"), Logical Hypothesis, and Statistical Hypothesis.

Here, suitable hypothesis testings are (Null Hypothesis (H_0), Alternate Hypothesis (H_1). For statistical measure, we should choose suitable α(Level of Significance) value and then have to collect data and determine probability associated with test statistics and compare it with a level of significance and make decisions based on its statistics.

As mentioned above if companies are importing more than 30% raw material new tax system will be introduced for them. If companies import less than30% raw material, we use the null hypothesis(H0) in case if they import more than 30% alternative hypothesis(H1) will be used. For testing the null hypothesis appropriate statistical testing should be selected. For example standard normal distribution(z).

$z = (p-\pi)/\sigma$, where $\sigma = \text{sqrt}(\pi(1-\pi)/n)$

Based on available data have to analyze either companies‘ raw material import 30% or more. Taking the above example into consideration 50 companies import a quarter of their raw material from other countries.

So here p = 25/50 = 0.5

$\sigma = \text{sqrt}((0.30)(0.70)/50)=0.0648$

here $Z = (p-\pi)/\sigma = (0.5–0.30)/(0.0648)=0.0370$

Now using standard normal distribution table, the probability of obtained z value of 0.0370 is 0.64431. i.e.$P(z \leq 0.0370)= 0.64431$.

Similarly, we have to calculate z value for the right side of the curve, we take assumed alpha value 0.05.

i.e rejected region is (1-0.64431) = 0.35569 and this probablitiy is comparable to alpha.

If Z is between 0.05(curves left value) and 1.65(curves right value)(normal probability is 0.95053)

Here observed value for test statistics is 0.35569; the p-value is 0.5 and π=0.30 is less than the level of significance.

The above-calculated Z values are 0.0370 is less than the value of 1.65.

Finally fail to reject H0 comparing with the level of significance, risk of committing towards false positives. Hence the null hypothesis is used and alternative hypothesis is rejected. So based on our evidence most of the companies import less than 0.30 raw materials.

Basically machine learning is a trial and error process because it is hard to find a perfect model that suits our data in our fist. Data validation is effective in nature but these statistical tests are designed to address the problem and calibrate the probability of the observed sample from the distribution. These statistical tests will help to develop suitable models for machine learning within a short amount of time and reduces resource utilization.

A. Data Validation in machine learning model:

Errors in the data affect the pipeline structure [19] that reflects the execution of the model output. Later finding and fixing those errors required to change the code which is tremendously hard to automate the pipeline process. In case, if we repair inconsistent data in order to correct the errors using probabilistic programs, but it immediately affects the model output.

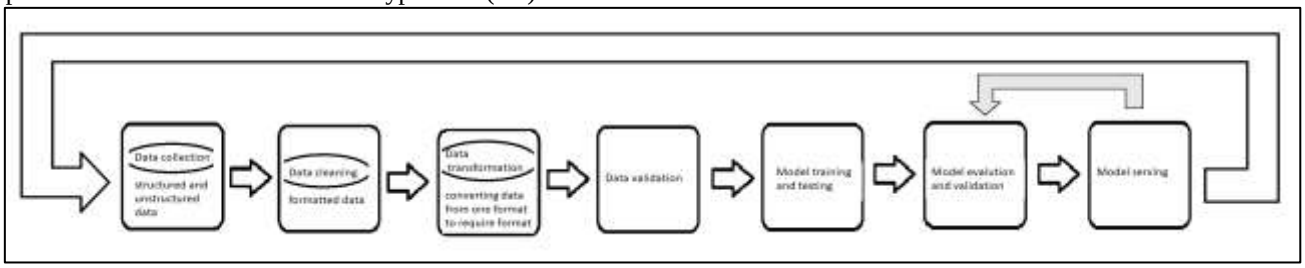

Fig. 1: Typical machine learning pipeline

1There are different kinds of data validation techniques are available such as Train/test split, k-Fold Cross-Validation, Nested Cross-Validation, Leave-one-group-out Cross-Validation, Leave-one-out Cross-Validation, 5x2CV paired t-test, McNemar’s test, Time-series Cross-Validation, Wilcoxon signed-rank test, 5x2CV combined F test. These data validation techniques help most of the time ensuring data integrity and usefulness.

So, in turn the data validation system must notify inconsistency in the data with probity, then its possible to solve data inconsistency problems before data passes to the next stage in the pipeline as shown in the Figure 1. Mostly ML engineers or data scientists can segregate the root cause of the problem.

We have to set a bandwidth for data stream in the pipeline, whenever model recognizes the inconsistent data, which is out of this bandwidth stream should notify the user about inconsistent data, and this data gets priority based on the effect it shows on the pipeline if this inconsistent data causes more biased output then, this data should be rectified immediately, otherwise it will damages end results. If inconsistent data does not cause any change in the pipeline process, it can be ignored.

The validation model have to control all the possible inconsistent data [10] that should not propagate down to the training model. If there is no new data introduced in the pipeline, the training and testing models code stays as earlier.

For example hotel booking web applications would like to predict future foreign customers density in country, based on current and past visited customers to a specific country from this application.

The key question to ask in this scenario which data properties affect significantly the model perfection and what dependencies affect the model infrastructure. After prioritizing the model hyperparameters and the behavior related to expected output results. Validating this input data is a high priority in the model, because this validated data passes to the training model. If there is any deviation between training and serving data, then this deviation causes abnormality in the serving scenario.

In order to control abnormality in the system finding possible deviations as early as possible in the system gives consistent output performance.

For example in our hotel booking web application some individual‘s details(about 10%) are missing in a new pipeline cycle, from which country they are visited and how old are they(age = 25 to 35), in this scenario data validation system should notify the user using probabilistic synopsis about missing features, details and user should take enough caution and correct missing values in order to avoid biased output.

These articulations allow users to understand better, the capacity of the errors, their ability on the model accuracy, and sensitivity of the notifications.

Detecting vulnerabilities [6] during validation increases model productivity because it reduces the financial loss for the company and prevent negative consequences in order to step back in the pipeline. Altogether data validation is a key element of ML infrastructure.

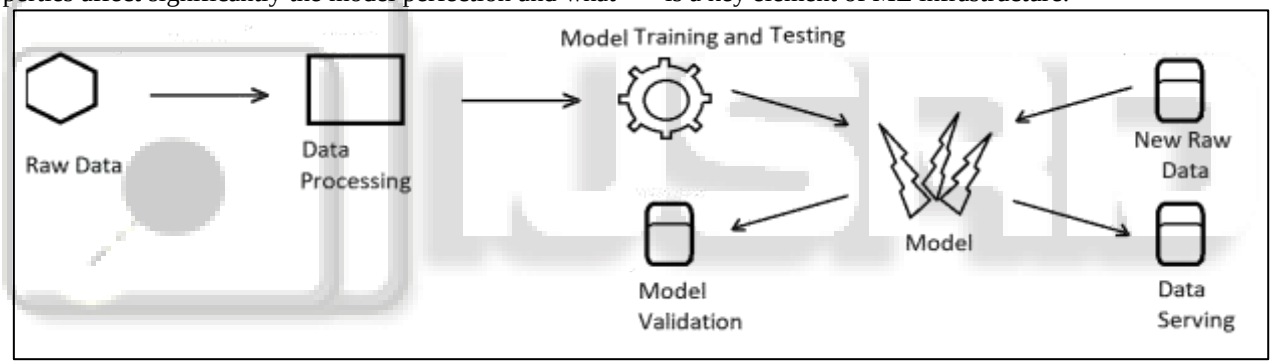

Fig. 2: Data validation process in Machine learning pipeline

ML completely depends on data, it requires to validate in order to perform well as shown in the Figure 2. It is most important to follow and understand features in the data what exactly they mean to the output in every pipeline cycle, for example in our hotel web application if we do not fallow case-sensitivity between two pipeline cycle, in the same manner, the pipeline takes data as different features rather than the same feature in both pipeline cycles. It's most important to take care of data sensitivity and format because it affects model accuracy.

In pipeline if we face new features after certain pipeline cycles, we have to create a new field, if this feature has priority and effect on output. In case after certain cycles in pipeline certain features will disappear from the model [9,16], we have to find origins for this irregularity in the data and rectifying these issues before passing data to the training model.

When identifying and correcting errors in the data there is a commutation between recall and precision, it will impact overall product output. Hence, it is important to balance the bandwidth between recall and precision based on feature importance and their effect on the result.

In the above hotel web application data arrives continuously in every cycle into the pipeline, as new data comes, it should not contain the same features as the previous cycle because the possibility to change the web application framework and it affects its features too. But the new pipeline cycle has more priority on new data compared to old data. And data validation should conduct prioritizing features in new data comparing with old data. Prioritizing features in new data is a hard task because it impacts the overall product.

Depending on the application and its service to the users and it requires an extensive amount of domain knowledge experts to prioritize these features in a data validation.

Selecting perfect data bandwidth most important because it has an impact on production.

So we have to track the errors and their impact after they fix in the pipeline is a big challenge because tracking model accuracy after fixed errors and analyzing their impact on the system is the most important criterion.

After data validation we have to feed this vectorized data into a training model for training, here the

model will be evaluated with different machine learning algorithms. After a model trained with algorithms, it will be deployed and served.

After serving data, the new pipeline cycle will start after a certain interval of time.

IV. FUTURE WORK

In this paper we presented a data validation system and how it works in a machine learning platform for better model accuracy and explained usual problems in data validation systems. This conceptual approach forms a basis for future research work on data validation because more work required on error detection, advanced analytics, and metrics. Required more attention to statistical techniques applied in the data validation system and another important field is to track the lifecycle of data and consumption of CPU or GPU usage and complete-time elapse for pipeline process.

V. CONCLUSION

Data validation is one of the main parts in the machine learning system. Providing clean and quality data is a major issue in the pipeline process but using proper data validation techniques as explained in the paper we can avoid biased machine learning models. This data validation system contains most of the common problems faced, while preparing data for model training. Finally, in this paper we encapsulated the data validation system and required infrastructure for preparing better data for machine learning systems.

REFERENCES

[1] N. Laranjeiro, S.N. Soydemir, J. Bernardino. A Survey on Data Quality: Classifying Poor Data. In Proc. of IEEE 21st Pacific Rim International Symposium on Dependable Computing (PRDC).
[2] B. Settles. Active Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool, 2012.
[3] M. Kim, T. Zimmermann, R. DeLine, and A. Begel. Data Scientists in Software Teams: State of the Art and Challenges. IEEE Transactions on Software Engineering, 2017.
[4] E. Rahm and H. H. Do. Data cleaning: Problems and current approaches. IEEE Data Eng.
[5] Tune H. Pers, Anders Albrechtsen, Claus Holst, Thorkild I. A. Sorensen and Thomas A. Gerds. The Validation and Assessment of Machine Learning: A Game of Prediction from High-Dimensional Data
[6] Jonathan Glowacki(FSA,CERA,MAAA), Martin Reichoff. (Milliman white paper). Effective model validation in machine learning.
[7] M. Yakout, A. K. Elmagarmid, J. Neville, M. Ouzzani, and I. F. Ilyas. Guided data repair. PVLDB, 2011.
[8] Mahesh. (towards data science). Everything You Need To Know about Hypothesis Testing
[9] Polyzotis, N., Roy, S., Whang, S. E., and Zinkevich, M. Data management challenges in production machine learning. In SIGMOD.
[10] Chervenak, A., Foster, I., Kesselman, C, Salisbury, C. and Tuecke, S. (2000). the data grid: Towards an architecture for the distributed management and analysis of large scientific datasets. Journal of Network and Computer Applications.
[11] Xu Chuy, Ihab F. Ilyasy, Christopher Ré. HoloClean: Holistic Data Repairs with Probabilistic Inference. (Stanford University and University of Waterloo)
[12] Eric Breck 1 Neoklis Polyzotis 1 Sudip Roy 1 Steven Euijong Whang 2 Martin Zinkevich .DATA VALIDATION FOR MACHINE LEARNING
[13] Chunli Xie, Jerry Gao, Chuanqi Tao. Big Data Validation Case Study
[14] Laura L. Pullum1,2, Chad Steed2, Sumit K. Jha3, Arvind Ramanathan2. Mathematically Rigorous Verification & Validation of Scientific Machine Learning
[15] Robert G. Sargent.(Syracuse University) Verification and validation of simulation models.
[16] Robert G. Sargent.(Syracuse University) Verification and validation of simulation models.
[17] Tobias Cagala. Improving Data Quality and Closing Data Gaps with Machine Learning
[18] Issam El Naqaa); Dan Ruan, Gilmer Valdes, Andre Dekker, Todd McNutt, Yaorong Ge, Q. Jackie Wu, Jung Hun Oh and Maria Thor, Wade Smith, Arvind Rao, Clifton Fuller, Ying Xiao, Frank Manion, Matthew Schipper, Charles Mayo, Jean M. Moran, and Randall Ten Haken. Machine learning and modeling: Data, validation, communication challenges.
[19] Dr. Christian Ruiz, Swiss Federal Statistical Office, Switzerland. Improving Data Validation using Machine Learning
[20] Breck, E., Cai, S., Nielsen, E., Salib, M., and Sculley, D. The ml test score: A rubric for ml production readiness and technical debt reduction. In IEEE Big Data.
[21] Ding, D. Zhang and X. H. Hu, "A Framework for Ensuring the Quality of a Big Data Service," IEEE International Conference on Services Computing (SCC)
[22] M. Stonebraker, D. Bruckner, I. F. Ilyas, G. Beskales, M. Cherniack, S. B. Zdonik, A. Pagan, and S. Xu. Data curation at scale: The data tamer system.
[23] K. H. Brodersen, F. Gallusser, J. Koehler, N. Remy, and S. L. Scott. Inferring causal impact using bayesian structural time-series models. Annals of Applied Statistics, 2015

