Journal of Multidisciplinary Engineering Science and Technology (JMEST)
ISSN: 2458-9403
Vol. 7 Issue 8, August - 2020

A Machine Learning Approach To Today's Metadata Management In BI, Data Analytics & Data Science Space

Pintu Kumar Ghosh
Standard Chartered Bank, Singapore

***Abstract*—As today's business is fully data driven and decision is based on the insights that have been gained through the data as available to the business. Data is available in heterogeneous formats that have been extracted from disparate sources with higher velocity and volumes. Big Corporates, IT, Business and any other industry leaders had to strategize as to how easily, efficiently and cost saving manner this can be achievable on time. Eventually there was ask to explore and expand business catalysts like Big Data, Open Source technologies, Algorithms, Data Structures, low cost Software and Hardware, right skillset to manage and take this forward. There has been a paradigm shift in Data Management, Data Governance, Master Data Management, Data Quality (DQ) and Metadata Management. We used to perform day-to-day business operation through traditional way of managing data and gaining insights with the help of Relational Database Management System (RDBMS), ETL(Extract, Transform and Load)/ELT(Extract, Load and Transform), EDW (Enterprise Data Warehouse), Business Intelligence (BI, Visualization, Reporting and Analytical Modelling tools and host of such types of tools. Thereby, Data Science has also become popular in particular Machine Learning techniques that has been delivering business value through large scale automation with faster pace. In this paper, I will try to explain how easily we could adapt [9] Machine Learning (ML) techniques to solve some of the problems arise through day-to-day data driven business operations. Right Data Quality Framework, Data Quality Management, Metadata Management and Data Governance practice is the key to success.**

***Keywords*—METADATA MANAGEMENT, DATA QUALITY, DATA ANALYTICS, DATA GOVERNANCE, DATA SCIENCE, MACHINE LEARNING. PREDICTIVE MODELLING, MODEL DEVELOPMENT, MODEL VALIDATION, MODEL DEPLOYMENT**

Introduction

Data is an asset to any organization. Nowadays, business decisions are taken through data and actionable insights. Managing data efficiently with automation and lower cost is the focus of top management. Even though data has been managed through best of the best possible way, how do we ensure that we maintain good data? If you maintain good data, how do we ensure that there is no repetition of data storage across your organization? If there is impact to your data as part of any business changes, how do you manage your data efficiently and cost saving manner? If all the above have been well managed and are robust enough, are we able to secure the data as per regulatory requirements? These asks are many folds. Now, we could focus on Data Quality related challenges. If we establish or follow appropriate policies and right algorithms to measure data problems and right data rules to detect anomalies and right skillset to analyse data and identify the root cause of the problems and manage thereafter for the strategic solutions. This was quite possible with the help of modern day's sophisticated Data Science technologies [12] like Machine Learning algorithms that can predict any data anomalies with accuracy without human intervention while finding out the tuples with anomalies or the features with outliers. Therefore, implementing predictive ML models to analyse data that were used for Analytical Models for predictive modelling is the key here. Metadata reports have been generated based on the above principles, reviewed and monitored ongoing basis as part of Data Governance practice. This approach have been helping or would help to save times and resources, in turn, it improves overall costs. This in turn would help the department to focus on their future data strategy and roadmaps. The paper has been written for the benefits of the Data Architect, Data Engineers, Data Scientist, Data Analysts, Data Architects, Project Managers and Data Stewardship professionals including Head of the Department (HOD) of Enterprise Information Management (EIM), Chief Data Office (CDO), Chief Information Office (CIO), Information Technology Office (ITO), CAO (Chief Data Analytics and Data Science Office) for Returns on Investment (ROI) and key strategy making processes.

Purpose and Benefits

Metadata Management, Data Quality Management and Data Governance is the key for the businesses driven by data. This is not possible to achieve the objectives or goals of Data Quality Management through mere data profiling techniques to find out data anomalies. The advanced techniques like AI (Artificial Intelligence), DL (Deep Learning), ML algorithms or so would be the show stoppers as they outperform [9] than human beings on specific tasks. It can predict data anomaly from the past data (Supervised learning

techniques) or no data at all (Unsupervised learning techniques) with automated approaches - without human interventions. In traditional methods, DQ Checks have been carried out through SQL like tool to validate the data based on list of data profiling rules which are predominantly columns and rows oriented unlike ML algorithms which are tuples and features oriented. Data anomalies have been detected through tuple investigations along with available features. The benefits are multi-folds in the sense that we could achieve cost saving and revenue earning objectives as compare to traditional approaches. Although there would be few challenges along the road like appropriate change management policies, organization silos, data quality issues, various stakeholders buy-ins, legacy systems, not utilizing the right toolsets, unavailability of the right resources or budget and so on. Once organization data has been enriched with high-quality there would be unlimited possibilities like Data Monetization [1], Data Democratization [2], and Data Commercialization [1] and so on. This, in turn, could open up new perspectives as if there is a dawn of a new revolution to your organization.

The approach – Evaluation of ML algorithms

Once the data has been collected, cleansed, enriched, integrated, processed and stored in Data Marts, EDW, ODS or so based on traditional way or just maintaining the data modern way [12] AS-IS (raw formats) in Data Lake or Data at Transit or Data at Rest (Figure – 8), there are asks from businesses to perform data completeness and accuracy checks in line with any regulatory (BCBS 239) and operational control requirements. In both the scenarios (in today's case: Figure – 1), a robust Data Quality Framework can be implemented where data anomaly can be detected and a strategic plan would be devised for data remediation.

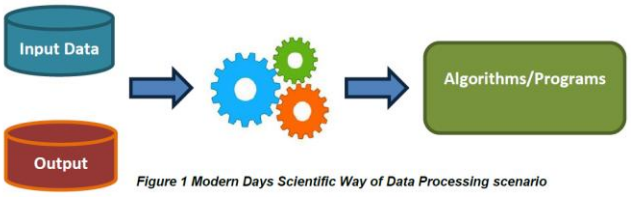

Figure 1 Modern Days Scientific Way of Data Processing scenario

Let us elaborate the current approach that has been depicted here - introduction to ML algorithms for predicting data outliers for BI, Data Analytics & Data Science projects.

✓ At first, identification of the right tools for AI/ML/DL algorithms out of various rich sets of tools that are available in the market.

✓ Check if there is any existing tool available to your organization else looking for assessments of on-boarding new tool.

✓ Check if this is cost effective and identifying ROI towards the final objectives.

✓ Ensure of on-boarding a tool for Data Quality assessments with AI/ML/DL capabilities.

✓ If the tool happens to be Open Source like Python, R, Matlab, Keras, Weka or so - are we ok to go ahead? If so, which are the AI/ML/DL algorithms should be considered for your projects?

✓ After final decision has been sought out, a detailed Project Plan would be charted out with the right resources and budgets outlined clearly.

✓ Find out if the list of assumptions/requirements have been documented in case there are missing values existed (Data Cleaning/Preparation) in the required data (Feature Engineering) and the possible data imputation techniques [8] without which prediction would not be accurate or the threshold of the outlier and inlier splits, the Training - Testing datasets splits [5], identification of optimized Hyper-parameters and so on.

✓ Ensure if the right ML features (Feature Engineering) [10] have been short listed as you would know ML (Scikit-Learn) does not support [3] categorical variables and selecting right label encoding or conversion techniques and documenting thereafter in order to carry out the implementation, if not, this would have already been selected for the existing Predictive Modelling exercise for BI (Business Intelligence), Data Analytics and Data Science projects.

✓ Check if processed data can be fitted to ML algorithm and also to perform feasibility study (Data Exploration) on the available data and also to ensure that data dimensionality reduction [11] techniques have been identified in case it would be used in the model and documenting thereafter.

✓ Check if there any Data-Ops, Dev-Ops, CI/CD, Orchestration, Micro-service toolsets are available and if not, ensure these toolsets have been identified for automated production deployment [13] or identification of manual deployment approach if need be.

✓ During implementation – start with the actual development of the ML models with the list of ML algorithms that were short listed for Data Anomaly [7] prediction exercise.

✓ Now, sample data has been prepared as Synthetic [19] or that have been taken from real world business with the right training [5] and testing splits for the candidate ML models.

✓ Keep following all the other steps of ML model building like cross validations [4], testing rigorously through few cycles until expected accuracy has been achieved.

✓ Finalize the ML model and deploy it in the production pipeline to predict outliers in the real data.

✓ Continue monitoring and validating the performance [20] of the ML models for predictive power, accuracy and stability and so on.

✓ Ensure model performance evaluation or validation[20] techniques such as Recall, Precision, F1 Score, Matthew's Correlation Coefficients, Confusion Matrix, k-fold cross-validation, ROC-AUC and so on scores shows up to the standard or as per your organization needs.

✓ If we could use LIME (Local Interpretable Model-agnostic Explanations) methods [9] that could easily explain to the internal and external regulators.

This section would depict various classification approaches - Supervised & Un-supervised learning whereby it was observed the expected accuracy in both the methods.

Training and Testing of ML Model

a. Supervised learning approach – Random Forest:

At first, the candidate model (with supervised learning [14]) was trained sample data of 80% with target (Class) variable with 1 or 0 whereby the outliers and inliers threshold kept at 0.0017. The same candidate model has been tested based on 20% sample with random_state=42 which is very popular parameter to use in ML algorithm. It was needed to perform basic data profiling, exploratory analysis and also found out the correlations of the 20+ features which were already transformed except Amount through ML compatible scale like Principal Component Analysis (PCA) [11] or so. Some of them (F2, F5) have shown correlations with Amount as shown in dark black colour. The below graph is quite significant to visualize features relationship. Here classification variable (Class) was already provided in the input dataset.

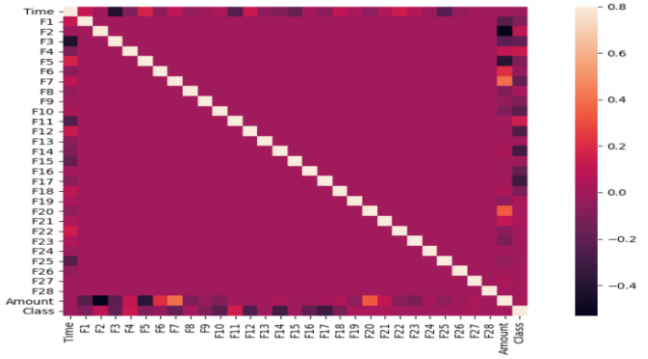

Figure 2 Correlation Matrix Generated by Python Sickit-Learn

Below are the results of our ML experiments:

i. Out of total observations (tuples – 284807 and 20+ features) examined in the sample dataset of financial industry.

ii. No of outliers(anomaly) in sample data: 492

iii. No of outliers (anomaly) **predicted** in test data (56,960) : 76

iv. No of error in test data: 23

v. The model used is Random Forest classifier - the accuracy is 0.9995962220427653

vi. The precision is 0.987012987012987

vii. The recall is 0.7755102040816326

viii. The F1-Score is 0.8685714285714285

ix. The Matthews correlation coefficient is 0.8747121626683524

There were series of ML algorithms have been scrutinized and evaluated. Results are shown as below.

ML Algorithm	No of errors	No of outliers/ anomalies	Accuracy	Precision	Recall	F1 Score	Mathew's Correlation Coefficients
Random Forest	**23**	**76**	**0.999596222**	**0.98701 2987**	**0.77551 0204**	**0.86857 1429**	**0.874712163**
Naive Bayes	398	62	**0.9930**12886	0.14622 6415	0.63265 3061	0.23754 7893	0.301961786
Logistic Regression	64	60	**0.9988**76444	0.69767 4419	0.61224 4898	0.65217 3913	0.653008313
Decision Tree - Gini	41	67	**0.9992**80222	0.87012 987	0.68367 3469	0.76571 4286	0.770946956
Decision Tree - Entropy	33	77	**0.9994**20666	0.86516 8539	0.78571 4286	0.82352 9412	0.824198194
Support Vector Machine(SVM)	88	29	**0.9984**5511	0.60416 6667	0.29591 8367	0.39726 0274	0.422165595

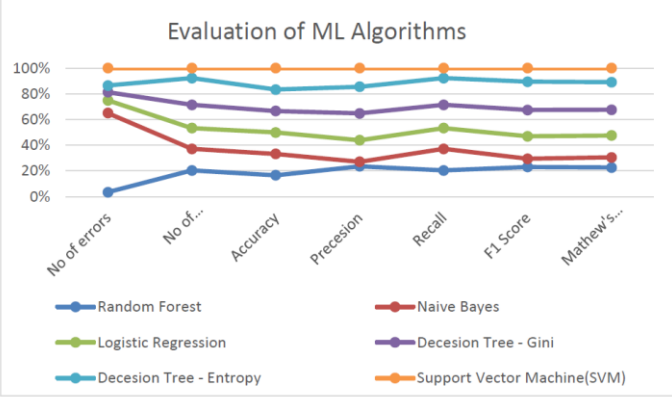

Figure 3 Comparison scores, matrices generated by Python Scikit-Learn

Based on Linear Regression (not recommended) approach, the score was 0.5166663660896537. Comparing the above, the accuracy is good to go ahead with "**Random Forest**" for the implementation if your use case is to use a supervised learning approach.

Confusion Matrix for Radom Forest:

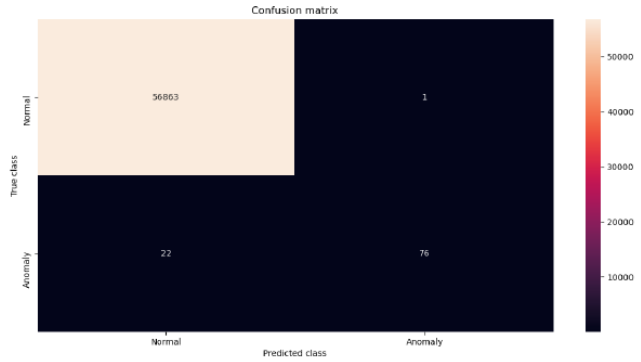

Figure 4 Confusion Matrix Generated by Python Scikit-Learn

Confusion Matrix is well established in order to evaluate ML model performance as described in the above through Recall and Precision scores. The results have also been evaluated through Accuracy F1 and Mathew's Correlation Coefficient scores.

b. Unsupervised learning approach – CBLOF:

An unsupervised learning [14] technique was in need to evaluate if there is any other algorithms that could outperform in anomaly/outlier detection with the above for the candidate model selection. Below are the list of ML algorithms that have undergone the evaluation process.

- ✓ Angle-based Outlier Detector (ABOD)
- ✓ **Cluster-based Local Outlier Factor (CBLOF)**
- ✓ Histogram-base Outlier Detection (HBOS)
- ✓ Isolation Forest
- ✓ K Nearest Neighbors (KNN)
- ✓ Average KNN

Upon using the same sample data as basis for the comparison, CBLOF outperformed. Predictions from the rest of the algorithm was far off. The only feature was selected here as Amount where we wanted to find out the anomaly. It was then transformed using MinMaxScaler approach. Below are the results.

i. Out of total observations (tuples – 284807 and 30 features) examined in the sample (test) dataset of financial industry taken from public source.
ii. No of outliers(anomaly) in test data: 492
iii. No of outliers(anomaly) **predicted** in test data: 489
iv. No of error in test data: 0
v. The model used is CBLOF classifier - the accuracy is 1.0
vi. The precision is 1.0
vii. The recall is 1.0
viii. The F1-Score is 1.0
ix. The Matthews correlation coefficient is 1.0

Comparing all the above algorithm under PYOD, the accuracy for CBLOF is good to go ahead with the implementation.

Prediction Plot:

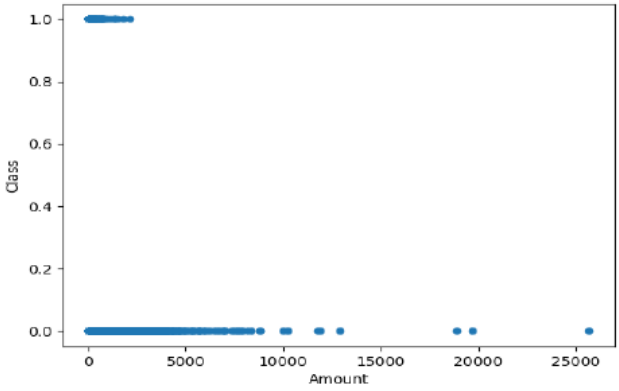

Figure 5 Prediction data point plotting Generated by Python Scikit-Learn

Confusion Matrix:

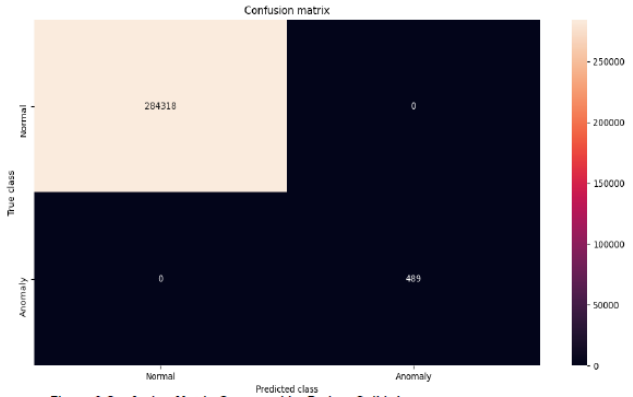

Figure 6 Confusion Matrix Generated by Python Scikit-Learn

It was concluded that Python PYOD [3] module based unsupervised algorithm outperforms than supervised algorithm – Radom Forest. This proves that the tuples with outliers were detected correctly for both the cases. However, either of them have been recommended for use based on your scenarios.

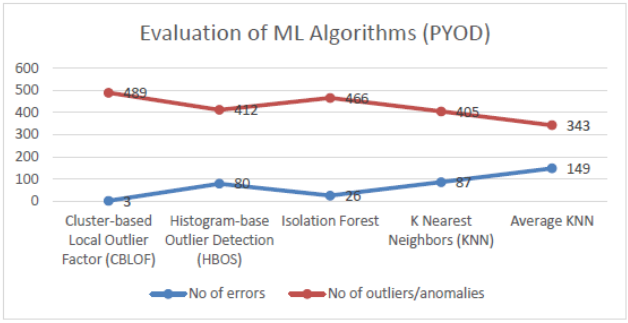

Figure 7 Comparison scores, matrices generated by Python Scikit-Learn

Implementation of ML Model

The data sample was taken above was collected was real-time indeed. After finalizing the above models, it was to be saved in the repository and to be deployed in the production pipelines. There are various approaches and best practices are available in this regards. In Python (Sickit-Learn) or any other tools can be used like Pycaret [13]. Most of the organization has been using state-of-the-art technologies for automated deployments [13] by making use of Dev-OPS or CI/CD or Data-Ops

techniques. There would be multiple models that were deployed and running in production. One should keep monitoring the performances of the models and fine-tuning hyper-parameter [21] would help to improve the accuracy, efficiency and productivity of the models. Depend upon the models and underlying data, use the right cross validation techniques to optimize the ML models for more accurate predictions. The scientific way of identifying outlies/anomalies in financial dataset was implemented as pilot study elsewhere with accuracy (100%) in perdition outcomes. Afterwards, these outliers check was carried out list of data elements required by business. Thereafter, these details were put together in a Data Quality report which is an important artefacts of Metadata Reporting and Data Governance.

Metadata Management and Reporting

Data Management, Metadata Management, Data Quality and Data Governance are tightly coupled. As we have learned there have been humongous amounts of data have generated on daily basis. Without appropriate strategy/policy like adapting to Machine Learning, AI practice, hard to achieve the revenue target and growth. The need of the hour is to devise a practice or cost-effective solution that will be help every organization at large scale without or minimal human supervision. Off late, there has been talk across regulatory firms that shifting traditional way of handling Predictive Modeling [16] to AI/ML/DL agnostic Predictive Modeling. Data Infrastructure is the key in order to support this initiative. Without good quality data and Metadata Management policies, it would be haywire or impossible. As such Data Quality checks can be leveraged through AI/ML/DL algorithms [18]. Based on the approach adopted here, the final ML model can be used for the identified data elements for profiling and exploratory analysis. Thanks to ML algorithms - anomaly detection, outlier checks, data imputation, fraudulent data identifications or so become possible. These types of metadata can be stored in tabular forms either in flat files or database or any other modern day's data formats. Thus, your organization could use such metadata for future data analysis and change impact analysis with the help of data lineage of the associated data elements. Data anomaly reports or dashboards will be presented further to the management or any internal or external body. Such reports, in turn, could be disseminated through leveraging any existing Data Visualization or BI/Reporting tools. There are various usage scenarios of data and associated metadata. Below is one of the example which have been visualized below while incorporating ML adaptation.

Data flow for a BI, Data Analytics & Data Science scenario:

Here is a day to day business operation scenario that generates lots of raw data and analysis has been accrued out either through processing or AS-IS format towards decision support systems like BI, Data Analytics & Data Science. The diagram (Figure - 8) would visualize the data flow of various data and their associated metadata. Sometimes, either EDW (Enterprise Data Warehouse) or Data Lake can be fit into your requirements and sometimes it does not. Hence, there would be requirement to devise a hybrid model for your data needs. In modern days, one of such data architecture can be similar to what has been represented as per below. In traditional way either through ETL (Extract, Load & Transform) /ELT process would be followed wherein in modern days, EL (Extract & Load) or ELT (Transform if need be) would be used with sophisticated tools/technologies like Data Ingestion, Batch Processing, Real-time Stream Processing or so.

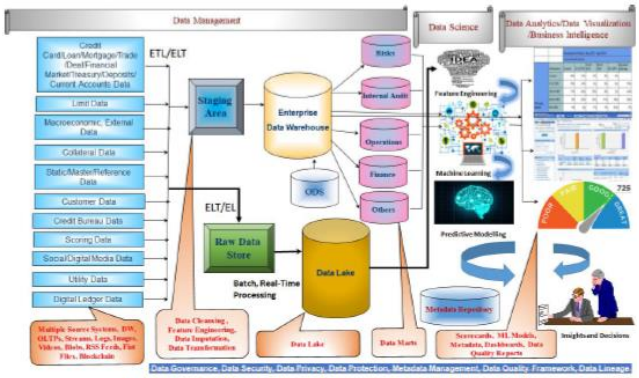

Figure 8 Modern Days Financial Sector Data Flow for BI, Data Analytics & Data Science scenario

Sample Metadata (outliers) can be shown below:

ID	Amount	Class
51	2.09	1
1388	2.69	1
1632	1	1
2453	1.29	1
2622	1.79	1
3780	126.21	1
7600	16.33	1
7778	14.49	1
9907	15.2	1
15364	45	1
15835	1602.36	1
17454	1	1
18471	527.24	1
18552	92.85	1
18570	6.28	1
18653	4	1
18738	2.68	1
19760	20.43	1
19920	0	1
20946	12.6	1
21440	105.55	1
22112	1	1

Result_CBLOF
489 OF 284807 RECORDS FOUND

ID	Amount	Predicti
1	149.62	1
565	43.34	1
3025	9.2	1
4640	79	1
5100	0.76	1
5756	0	1
6054	68.9	1
6368	36	1
7011	0.01	1
7846	4.9	1
8456	9.99	1
9203	169	1
10634	1.98	1
11066	11.85	1
16473	24.99	1
16564	8.95	1
16592	6.25	1
16835	14.34	1
17345	10	1
18130	40	1
18544	220	1
18785	156	1

Result_Random Forest
76 OF 56960 RECORDS FOUND

Below is the sample Data Quality (Metadata) report:

Data Source	Datamart	Dataset Name	Feature Names	Total No Tuples	No of exceptions	%	Materiality	Data Quality Rule	Data Domain Owner
EDW	Creditcard	Credit_Details	Amount	284,897	489	0.17%	45,207.07	Data **anomaly** found	Risk, Operation
EDW	Creditcard	Credit_Details	Amount	284,897	0	0.0%	0	Missing Values	Risk, Operation
EDW	Creditcard	Credit_Details	ID	284,897	0	0.0%	0	Missing Values	Risk, Operation
EDW	Creditcard	Credit_Details	Month	284,897	0	0.0%	0	Invalid Values	Risk, Operation
EDW	Creditcard	Credit_Details	No_of_cards	284,897	765	0.27%	3045.5	Missing values	Risk, Operation
EDW	Creditcard	Credit_Details	No_of_cards	284,897	37	0.01%	9827.31	Range of values	Risk, Operation

The metadata associated to Data Protection, Data Security, Data Privacy and Cyber Security [17] can also be identified and managed by leveraging ML algorithms through SPAM Email Filtering, Intrusion Detection, Malware Detection, Sentiment Analysis, Document Classification and Vulnerability Detection and so on. The metadata should be extracted through modern day's toolsets from underlying data that can be simple Logs, Texts, Streams, Images, Videos and so on. Afterwards, the same Machine Learning modelling lifecycle would be followed to predict the identified classification (2-way or multi-class) variables accurately on the aspects as mentioned in this section. Below are the sample metadata reports that would be leveraged through ML algorithm.

Data Source	Datamart	Dataset Name	Feature Names	Total No Tuples*	No of cases found*	Data Domain Owner	Associated Metadata	Point of Contacts
Data Lake	Legal	Legal_Dcouments	Document_ID	940,843	67	Legal, HR	Document is sensitive, restricted, public or not	<Name of the person in charge from each data domain owner>
Data Lake	IT Risks	Cyber_Data	Data_Accessed	12,589,745	34	IT, Risk	Intrusion detected or not	
Data Lake	Internal Audit	Application_Logs	Log_Details	73,283,489	231	IT, Risk, Audit	IP address/Password shown in the log or not	
Data Lake	IT Risks	Surveillance	Video_ID	12,986,572	566	IT, Risk	Sensitive word used in any video or not	
Data Lake	Compliance	Transaction	Recepient_Name	3,114,456,645	241	Compliance, Risk	Backlisted/Sanctioned or not	
EDW	Finance	Balance_Sheet	Exposure	221,214,536	21	Finance, Risk	Data is secured through access matrix or not (tuple or dataset level)	

* These numbers are strictly samples and demonstration purpose only

Likewise every department with their own BI, Data Analytics & Data Science requirements could be fulfilled with the similar approaches. Once appropriate data infrastructure is available, with appropriate Data Governance polices, Data Quality management frameworks and with right data and metadata management principles, success would be achieved. These metadata reports would guide them to adopt vision, strategy and roadmaps.

Post-Implementation and Way Forward

Similar approaches can be used and referred elsewhere irrespective industry or sectors. As we all know that ML model parameters would be tuned/optimized further based on the data available to you. The quality of the ML parameters would vary datasets to datasets. The list of ML algorithms adopted here was fit for the purpose. It could vary to your use cases be it for anomaly detection or outlier detection or credit scoring or churn prediction or spam detection or any other prediction exercises and which is also depending upon the underlying data. There would be some other sophisticated model validation [20] like K-fold cross-validation, ROC - AUC, Kappa or such techniques which are out of scope of this paper. There are regulators who would have been started encouraging such approaches [16] as this would be the way forward. ML algorithms are there in the market for quite sometimes now. The adoptability is the key here as these can be trusted and interpreted easily. Most of the organizations have started moving towards the same direction. Methodologies around Data Cleaning, Data Preparation, Data Transformation, Data Integration, Data Ingestion and Metadata Reporting have been there long time in the

market. The features have been adopted here as primary objective is to showcase the capability, adoptability and benefits of the approaches. Initially this could be used as pilot approach and then deploy it at large scale.

Conclusion

There have been numerous uses cases for this approach adopted here. Data can generated through various disparate sources, without identifications of the data or without identifications of the right data, it may not be relevant to the business. Subsequently, there was requirement to identify the metadata for the available data. Once your organization has been identified metadata and data lineage, accuracy and completeness to be maintained. Upon collecting and maintaining the right data and associated metadata, AI/ML/DL algorithms would be useful in order to resolve Data Quality issue [18] through automated and early-detection process. This would yield added values to the business with quick turnaround time. The idea or approach depicted here has also been validated where almost the same accuracy was observed. This paper would recommend to continue using automated approaches through AI/ML/DL that have been scientifically proven solutions. Finally, it would continue to contribute to your organization in multifold manners like Data Monetization, Data Democratization and Data Modernization, Data Commercialization and so on. As the time passes, Automation Testing & Back-Testing of existing production models would be required. This, in turn, would lead to set up new infrastructure to support this. As there has been growing needs of data[15] in the way of implementing Predicting Modelling extending existing data requirements to Blockchain platforms [15], Social Media or so. Thereby, there has been growing Data Protection, Data Security, Data Privacy and Cyber Security [17] requirements. The associated metadata are also critical to your organization and ought to be managed appropriately. Thus, there are requirements of the policies like the one that has been depicted here for the right Data Governance, Data Management, Data Quality Management, Metadata Management and Data Lineage in order to support BI, Data Analytics & Data Science practices. Hope this approach would help to your organization to match the growing needs.

Acknowledgements

I would like to thank to them whoever supported me directly or indirectly whenever required.

Contact Information

Your comments and questions are valued and encouraged. Contact the author at:

Author: Pintu Kumar Ghosh

Place: Singapore

E-mail: ghosh.pintu@gmail.com

Any brand and product names used here are trademarks of their respective companies or open source licensing.

References

1. Data Monetization, Commercialization background can be found at

https://sloanreview.mit.edu/article/demystifying-data-monetization/

https://en.wikipedia.org/wiki/Data_monetization

Llewellyn D W Thomas Dr Aija Leiponen “Big data commercialization” -

IEEE Engineering Management Review 44(2) · December 2016

https://www.researchgate.net/publication/304674108_Big_data_commercialization

2. Data Democratization background can found at

https://tdan.com/growing-importance-of-data-democratization/22138

https://heap.io/blog/data-stories/how-to-build-a-data-democratization-strategy

3. Why Python with Sickit-Learn & Big Data? Available at

https://www.whizlabs.com/blog/python-and-big-data/

https://scikit-learn.org/stable/auto_examples/index.html#decomposition-examples

https://docs.python.org/3.8/

http://wiki.apache.org/hadoop

https://towardsdatascience.com/why-is-python-programming-a-perfect-fit-for-big-data-5ac54ee8f95e

https://pyod.readthedocs.io/en/latest/

Kedar Potdar Taher S. Pardawala Chinmay D. Pai “A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers” International Journal of Computer Applications (0975 – 8887) Volume 175 – No.4, October 2017 https://www.researchgate.net/publication/320465713_A_Comparative_Study_of_Categorical_Variable_Encoding_Techniques_for_Neural_Network_Classifiers

4. Machine Learning model selection approaches are available at

Sebastian Raschka “Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning” arXiv:1811.12808v1 [cs.LG] 13 Nov 2018

https://sebastianraschka.com/pdf/manuscripts/model-eval.pdf

5. Machine Learning Training data requirements

Marius Muja David G. Lowe Scalable “Nearest Neighbor Algorithms for High Dimensional Data” - IEEE Transactions on Pattern Analysis and Machine Intelligence (Volume: 36 , Issue: 11 , Nov. 1 2014)

https://ieeexplore.ieee.org/document/6809191

6. Machine Learning algorithm requirements described at

O. Obulesu ; M. Mahendra ; M. ThrilokReddy “Machine Learning Techniques and Tools: A Survey” - 2018 International Conference on Inventive Research in Computing Applications (ICIRCA) July 2018

https://ieeexplore.ieee.org/document/8597302

Dominique Guégan Bertrand Hassani Regulatory learning: How to supervise machine learning models? An application to credit scoring” - The Journal of Finance and Data Science Volume 4, Issue 3, September 2018, Pages 157-171

https://www.sciencedirect.com/science/article/pii/S2405918817300648

Anastasios Petropoulos, Vasilis Siakoulis, Evaggelos Stavroulakis and Aristotelis Klamargias “A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting” - Ninth IFC Conference on “Are post-crisis statistical initiatives completed?” Basel, 30-31 August 2018

https://www.bis.org/ifc/publ/ifcb49_49.pdf

Dumitrescu, Elena_, Hue, Sullivany, Hurlin, Christophez, Tokpavi Sessi “Machine Learning for Credit Scoring: Improving Logistic Regression with Non Linear Decision Tree Effects” February, 2018

https://www.researchgate.net/publication/318661593_Machine_Learning_for_Credit_Scoring_Improving_Logistic_Regression_with_Non_Linear_Decision_Tree_Effects

7. Data Anomaly detection rationale can be found at

Dina Sukhobok ; Nikolay Nikolov ; Dumitru Roman “Tabular Data Anomaly Patterns” - 2017 International Conference on Big Data Innovations and Applications (Innovate-Data) Aug. 2017

https://ieeexplore.ieee.org/document/8316296

İlker Kalaycı ; Tuncay Ercan “Anomaly Detection in Wireless Sensor Networks Data by Using Histogram Based Outlier Score Method” - 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) Oct. 2018

https://ieeexplore.ieee.org/document/8567262

8. Missing values impacts on Predictions and ML classifications and imputation requirements

i. Talayeh Razzaghi,Oleg Roderick,Ilya Safro ,Nicholas Marko “Multilevel Weighted Support Vector Machine for Classification on Healthcare Data with Missing Values” Plos One May 2016

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0155119

ii. Alireza Farhangfar Lukasz Kurgan Jennifer Dy “Impact of imputation of missing values on classification error for discrete data” ELSEVIER Pattern Recognition - Volume 41, Issue 12, December 2008, Pages 3692-3705

https://www.sciencedirect.com/science/article/abs/pii/S003132030800201X?via%3Dihub

iii. A. Rogier T. Dondersa,b,c, *, Geert J.M.G. van der Heijdenc , Theo Stijnend , Karel G.M. Moons “SPECIAL SERIES: MISSING DATA Review: A gentle introduction to imputation of missing values” - Journal of Clinical Epidemiology 59 (2006) 1087e1091 http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.578.5816&rep=rep1&type=pdf

iv. TAPIO SCHNEIDER “Analysis of Incomplete Climate Data: Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values” *J. Climate* (2001) 14 (5): 853–871.

https://journals.ametsoc.org/jcli/article/14/5/853/29100/Analysis-of-Incomplete-Climate-Data-Estimation-of

9. Why to adopt and trust Machine Learning prediction? Can this be explainable?

i. Trevor Hastie Robert Tibshirani Jerome Friedman “The Elements of Statistical Learning Data Mining, Inference, and Prediction” Book - Second Edition
https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12.pdf

ii. Marco Tulio Ribeiro Sameer Singh Carlos Guestrin “Why Should I Trust You? Explaining the Predictions of Any Classifier” – KDD 2016

https://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf

iii. https://www.oreilly.com/radar/ideas-on-interpreting-machine-learning/

iv. Zachary C. Lipton “The Mythos of Model Interpretability” - arXiv:1606.03490v3 [cs.LG] 6 Mar 2017

https://arxiv.org/pdf/1606.03490.pdf

v. Finale Doshi-Velez Been Kim “Towards A Rigorous Science of Interpretable Machine Learning” - arXiv:1702.08608v2 [stat.ML] 2 Mar 2017

https://arxiv.org/pdf/1702.08608.pdf

10. ML Feature selection approach

Brian Barr Ke Xu Claudio Silva Enrico Bertini Robert Reilly C. Bayan Bruss Jason D. Wittenbach “Towards Ground Truth Explainability on Tabular Data” arXiv:2007.10532v1 [cs.LG] 20 Jul 2020

https://arxiv.org/pdf/2007.10532.pdf

Yoga Pristyanto ; Sumarni Adi ; Andi Sunyoto "The Effect of Feature Selection on Classification Algorithms in Credit Approval" - 2019 International Conference on Information and Communications Technology (ICOIACT) Jul 2019 https://ieeexplore.ieee.org/document/8938523

11. ML Dimension Reduction approaches described at

Yu Liang ,Arin Chaudhuri , and Haoyu Wang "VISUALIZING THE FINER CLUSTER STRUCTURE OF LARGE-SCALE AND HIGH-DIMENSIONAL DATA" arXiv:2007.08711v1 [stat.ML] 17 Jul 2020

https://arxiv.org/pdf/2007.08711.pdf

12. Transition to Modern BI and Data Analytics can be found at

https://www.eckerson.com/articles/ten-characteristics-of-a-modern-data-architecture

https://www.atscale.com/blog/the-six-modern-principles-of-modern-data-architecture/

https://www.snowflake.com/blog/beyond-modern-data-architecture/

https://www.forbes.com/sites/forbestechcouncil/2019/04/05/why-the-modern-day-corporation-should-consider-a-data-estate/#3c875ecc63ba

13. ML Model deployment approach referred here

https://pycaret.org/create-model/

José M. Alves ; Leonardo M. Honório ; Miriam A. M. Capretz "ML4IoT: A Framework to Orchestrate Machine Learning Workflows on Internet of Things Data" IEEE Access (Volume: 7) Page(s): 152953 - 152967

https://ieeexplore.ieee.org/document/8876834

Nicolas Ferry ; Phu H. Nguyen "Towards Model-Based Continuous Deployment of Secure IoT Systems" - 2019 ACM/IEEE 22nd International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C) Sep 2019

https://ieeexplore.ieee.org/document/8904644

Zheyi Chen ; Tao Xiang ; Xing Chen "Model-Driven Approach to Hadoop Deployment in Cloud" - 2017 5th IEEE International Conference on Mobile Cloud Computing, Services, and Engineering (MobileCloud) Jun2017

https://ieeexplore.ieee.org/document/7944885

Oscar Castro-Lopez ; Ines F. Vega-Lopez "Multi-target Compiler for the Deployment of Machine Learning Models" - 2019 IEEE/ACM International Symposium on Code Generation and Optimization (CGO) Mar 2019

https://ieeexplore.ieee.org/document/8661199

14. Supervised & Unsupervised learning defined here

https://en.wikipedia.org/wiki/Unsupervised_learning

https://en.wikipedia.org/wiki/Supervised_learning

15. Data Requirement Trends as shown by Gartner

https://pages.dataiku.com/2020-ai-trends

https://pages.dataiku.com/gartner-top-10-trends-data-analytics-2020

"What Is a Blockchain?" MIT Technology Review April 2018 https://www.technologyreview.com/2018/04/23/143477/explainer-what-is-a-blockchain/

16. Regulator reviews on Data Analytics and Data Science for Finance Sector

Monetary Authority of Singapore "Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector"

https://www.mas.gov.sg/~/media/MAS/News%20and%20Publications/Monographs%20and%20Information%20Papers/FEAT%20Principles%20Final.pdf

17. Data Security and Privacy related information

European Union. 2016. General Data Protection Regulation, Reg. (EU) 2016/679.

https://gdpr-info. eu/

https://www.csa.gov.sg/gosafeonline/resources/cybersecurity-and-data-protection-ecosystem

https://www.pdpc.gov.sg/Overview-of-PDPA/The-Legislation/Personal-Data-Protection-Act

https://securityboulevard.com/2018/04/20-important-data-privacy-questions-you-should-be-asking-now/

18. Data Quality requirements for ML projects as stated below

Anna Karanika Panagiotis Oikonomou Kostas Kolomvatsos Christos Anagnostopoulos "On the Use of Interpretable Machine Learning for the Management of Data Quality" - arXiv:2007.14677v1 [cs.LG] 29 Jul 2020

https://arxiv.org/pdf/2007.14677.pdf

19. Why it is important synthetic data for Machine Learning models

Yang Yue ; Ying Li ; Kexin Yi ; Zhonghai Wu "Synthetic Data Approach for Classification and Regression" - 2018 IEEE 29th International Conference on Application-specific Systems, Architectures and Processors (ASAP) Jul 2018

https://ieeexplore.ieee.org/document/8445094

Runzong Liu ; Bin Fang ; Yuan Yan Tang ; Patrick P.K. Chan “Synthetic Data Generator for Classification Rules Learning” - 2016 7th International Conference on Cloud Computing and Big Data (CCBD) Nov 2016

https://ieeexplore.ieee.org/document/7979934

Chao Tan “A Model-Based Approach to Generate Dynamic Synthetic Test Data” - 2019 12th IEEE Conference on Software Testing, Validation and Verification (ICST) Apr 2019

https://ieeexplore.ieee.org/document/8730199

20. ML algorithm performance evaluation has been explained here

Janelyn A. Talingdan “Performance Comparison of Different Classification Algorithms for Household Poverty Classification” - 2019 4th International Conference on Information Systems Engineering (ICISE) May 2019

https://ieeexplore.ieee.org/document/8954587

Fubao Zhu ; Xiaonan Li ; Daniel Mcgonigle ; Haipeng Tang ; Zhuo He ; Chaoyang Zhang ; Guang-Uei Hung “Analyze Informant-Based Questionnaire for The Early Diagnosis of Senile Dementia Using Deep Learning” - IEEE Journal of Translational Engineering in Health and Medicine (Volume: 8) Dec 2019

https://ieeexplore.ieee.org/document/8933438

21. Fine-tune Hyper-parameters in ML models stated as

Binghui Wang ; Neil Zhenqiang Gong “Stealing Hyperparameters in Machine Learning” - 2018 IEEE Symposium on Security and Privacy (SP) – May 2018

https://ieeexplore.ieee.org/document/8418595

Zhen Wang ; Mulya Agung ; Ryusuke Egawa ; Reiji Suda ; Hiroyuki Takizawa “Automatic Hyperparameter Tuning of Machine Learning Models under Time Constraints” - 2018 IEEE International Conference on Big Data (Big Data) Dec 2018

https://ieeexplore.ieee.org/document/8622384

Klára Peškova ; Roman Neruda “Hyperparameters Search Methods for Machine Learning Linear Workflows” - 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA) Dec 2019

https://ieeexplore.ieee.org/document/8999298

Tinu Theckel Joy ; Santu Rana ; Sunil Gupta ; Svetha Venkatesh “Hyperparameter tuning for big data using Bayesian optimisation” - 2016 23rd International Conference on Pattern Recognition (ICPR) Dec 2016

https://ieeexplore.ieee.org/document/7900023