Optimizing Dataset Quality for Enhanced Machine Learning Performance

A Study on the Impact of Dataset Metrics

Efe Unluyurt
Supervisor(s): Kubilay Atasu, Atahan Akyildiz
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Efe Unluyurt
Final project course: CSE3000 Research Project
Thesis committee: Kubilay Atasu, Atahan Akyildiz, Burcu Kulahcioglu Ozkan

1 Abstract

With the increase of machine learning applications in our every-day life, high-quality datasets are becoming necessary to train accurate and reliable models. This research delves into the factors that contribute to a high quality dataset and examines how different dataset metrics affect the performance of machine learning models particularly focusing on Graph Neural Networks (GNNs) Tabular Transformers and Large Language Models (LLMs). The metrics, under scrutiny include graph sparsity, missing data cells, modularity and text length. Various datasets are adjusted to assess how these metrics impact model performance.

The results of the experiments reveal that sparse graphs can preserve relational information. However increasing density does not necessarily lead to improved performance due to noise interference. The models demonstrated accuracy and low error rates in the presence of significant missing data indicating their ability to handle incomplete information effectively and generalize well based on imputation strategies and structural design. Higher modularity was found to aid in capturing patterns. Introduced complexity that could potentially hinder performance. Notably text length emerged as a factor influencing model performance by offering contextual details.

These insights show the significance of considering attributes when designing machine learning models for intricate predictive tasks. Through experimentation and optimization of these metrics we can enhance model resilience and accuracy for applicability, in real world scenarios.

2 Introduction

The quality and characteristics of datasets play an important role in the performance of machine learning models, especially where Graph Neural Networks (GNNs), Large Language Models (LLMs), and Transformers are used. This paper aims to lay a foundation about the important aspects that make a dataset great. Understanding and improving dataset quality when performing machine learning tasks is important because the data's integrity directly influences the performance and reliability of the models [14].

The main research question of this paper is: What defines a high-quality dataset, and which metrics most accurately assess its usefulness? To address this question, the paper aims to explore a set of dataset metrics. These metrics will provide an outline about how they can contribute to a better dataset used in machine learning tasks.

Throughout the project, we will be introducing the datasets such as OGBN Arxiv (Open Graph Benchmark) [9], Amazon-Fashion [10], and the IBM Transactions for Anti Money Laundering dataset [2]. They will then be modified and structured to be usable for specific machine learning tasks. This paper aims to analyse these datasets to understand their structural and statistical nuances, and then use the datasets to see which metrics contribute to the quality of a dataset.

Through the course of this paper, we will explore the methodology for creating and evaluating these datasets, detail the experimental setups including the sources and types of data used. With that, this paper aims to give an answer to the main question, "What makes a great dataset?". The conclusion will reflect on the impact of the metrics and propose directions for future research that could further refine dataset construction and evaluation in the field of machine learning.

Finally, this introduction outlines the scope and ambition of the project, setting the stage for an in-depth exploration of dataset quality and construction, aiming to significantly enhance the tools available to data scientists and researchers in the field.

3 Background Information

In the machine learning field, it is known that the quality of datasets is important for creating great models.[14] As machine learning models, especially those employing Graph Neural Networks (GNNs), Large Language Models (LLMs), and Transformers, become more advanced, the datasets they are trained on must be carefully constructed and evaluated to ensure the reliability of the models. This paper aims to go deep into the characteristics that define a high-quality dataset, particularly focusing on their application in machine learning tasks.

3.1 Importance of Dataset Quality

The integrity and quality of datasets directly influence the performance of machine learning models. Poor-quality data can lead to inaccurate models, which can significantly affect the outcomes in applications such as fraud detection and financial analysis. Hence, understanding the factors that contribute to dataset quality is important. Each of these factors can impact the model's ability to learn and generalize from the data, affecting its predictive performance. Therefore, it is important to learn about these features.[13].

3.2 Historical Context: Data in Machine Learning

The importance of data in machine learning was recognized early on. In the 1990s, the UCI Machine Learning Repository was established, providing a resource for researchers. Over time, the scale and complexity of datasets have increased, leading to the development of benchmarks such as the MNIST dataset for handwritten digit recognition and the ImageNet dataset, which spurred advancements in computer vision [11, 4].

3.3 The Role of Graphs in Machine Learning

Graphs have become increasingly important in machine learning, particularly with the rise of Graph Neural Networks (GNNs). GNNs are designed to work with graph-structured data, capturing the relationships between entities. The Open Graph Benchmark (OGB) is one such initiative aimed at providing standardized datasets for evaluating GNNs [8].

3.4 Previous Research and Literature

Previous research has explored various aspects of dataset construction, evaluation and augmentation. For example, the "Open Graph Benchmark: Datasets for Machine Learning on Graphs" focuses on creating datasets optimized for graph-based machine learning tasks. However, the literature often lacks detailed analysis on the suitability of these datasets for specific tasks, such as fraud detection or financial crime analysis. In addition there are other papers like the "A survey on dataset quality in machine learning" [1], which explores the importance and challenges of dataset quality in machine learning applications. However, they do not necessarily explore what makes datasets great, but they focus on the importance of the quality of them instead. This paper aims to fill this gap by conducting an in-depth exploration of different datasets and their applicability to various machine learning tasks [12]. This way, the final goal is to find out about the dataset metrics that make a great dataset for a specific task.

3.5 Contribution to the Field

By focusing on the detailed analysis of various datasets, this paper aims to contribute to the field of machine learning. It aims to provide an understanding of some of the factors that make a great dataset. This, in turn, can lead to the development of more accurate and reliable machine learning models.

4 Methodology

This chapter outlines the methodology defined to address the research questions, which focuses on obtaining and formatting datasets, experimenting with different metrics, and evaluating their impact on various machine learning models. The objective is to identify the metrics that differentiate datasets and determine which metrics make a dataset better suited for specific models and why.

4.1 Data Acquisition

The initial step was to collect datasets that can be used with the models that we have, which can provide diverse types of data relevant to the research objectives. The datasets included OGB (Open Graph Benchmark)[9], which offers a variety of graph datasets for machine learning research; Amazon-Fashion[10], focusing on reviews on Amazon fashion products; IBM AML[2], which contains fraud and non-fraud banking transactions; and Ethereum Phising Transaction Network [18], which includes data about Ethereum transactions.

These datasets are introduced and prepared to evaluate the performance of different models that the group is working on. This is why different kinds of datasets are chosen, since only some datasets can work with specific models.

Dataset	**Type**	**Data Content**
Amazon-Fashion	Text-based, graph structured	Amazon customer reviews and ratings on fashion products
IBM-AML	Numeric, graph structured	Synthetic transaction records with flags indicating suspicious transactions.
Ogbn-arxiv	Numeric, graph structured	A citation network of Computer Science (CS) papers from arXiv. (Text values are represented as node2vec vectors)
Ogbn-arxiv (text)	Text-based, graph structured	A citation network of Computer Science (CS) papers from arXiv.
Ethereum Phishing Transaction Network	Numeric, graph structured	Transaction records related to Ethereum phishing activities

Table 3.1: Description of the datasets considered

4.2 Data Preparation

To be able to use the datasets with models we have, it is first needed to apply pre-processing step (if needed), and convert it to a PyTorch[16] Dataset object. Tabular data were converted into CSV format and structured to include the necesarry features, while graph data were represented in formats compatible with graph neural network (GNN) libraries, and also in a tabular format, which is necessary for testing the dataset with other models.

4.3 Procedure

Initial experiments starts with the analysis of each graph. There are several metrics that could change the model's performance. The metrics that will be used in this project will be shown in Section 5. After analysing each dataset, finding the differences between the datasets is crucial to learn about which datasets can be used for testing out which metric. At the end, the goal is to explore if there is a correlation between these metrics and model performance to identify patterns and insights.

5 Experimental Setup

To evaluate the dataset quality, several key metrics have been chosen. These metrics will be run on the model for comparison to determine if we can conclude that they contribute to a great dataset and improve the model's performance.

Metric	**Importance**
Graph Sparsity	Some models may perform better on sparse graphs due to it's simple structure. Knowing the sparsity can help in choosing and tuning these algorithms for optimal performance.
Number of missing cells	A high number of missing cells can indicate poor data quality, which can adversely affect the performance of machine learning models.
Modularity	Identifying the structure of communities might reveal insights into the natural divisions within the dataset.

Table 4.1: Importance of the general metrics.

However, different metrics are also used to evaluate different types of datasets, as an example, since we are using LLMs, some other metrics such as the length of texts could be analysed for text heavy datasets, since it could also effect the performance of LLMs.

Metric	**Importance**
Text Length	The length of the text could influence LLM's behaviour, this can cause the model work better or worse.

Table 4.2: Importance of the metrics for text-heavy datasets

Each metric requires it's own analysis, therefore it is not possible to use the same results to compare each metric. To overcome this, it is necessary to conduct different experiments to assess each metric.

To evaluate the impact of the graph sparsity on the model performance, we are going to use the IBM dataset, since it has initial sparsity of 0.99 and it is relatively a large dataset, therefore reducing the sparsity won't change the graph structure like the others. The approach is to use the same dataset, but modifying it to have different versions of it with different sparsity rates. This way, the same model will be run on all the modified datasets, and the results will help us to assess the impact.

The datasets we have don't have any missing cells, but to assess the impact of number off missing cells, the IBM dataset will be used, since it hasmany columns and many rows that missing values can occur, therefore it is possible to insert meaningful and realistic missing cells. The dataset will be modified to include missing cells, and again the same model will be run on all the modified datasets, to help us assess the impact of the number of missing cells in a dataset on the quality of it.

To assess the impact of modularity, the IBM-AML dataset will be used, since it has many communities already and better defined than other datasets. It will also be modified in a way that some of the edges between the communities will be removed, to change it's structure and change the modularity. This will be used to help us assess the impact ofmodularity of a dataset on the quality of it.

For the text length metric, the Amazon-fashion dataset will be used. The dataset will be splitted up into different batches, with different average length text. The model will be trained and evaluated on this different batches, too see how the model works with datasets that include different text lengths.This experiment is important since the text length can directly influence the LLMs performance, which also influences the model's performance

6 Results

To take the results for each metric, only 2 different models used. The first one is a model which contains a backbone, which consists of integrated Graph Neural Networks (GNNs) and Tabular Transformers. The GNNs handle data with network structures, capturing relationships between nodes, while the Tabular Transformers process structured, tabular data to identify patterns and dependencies. These two components work together to fuse the information, transforming the raw input embeddings. These embeddings, are then fed into decoders which is specialized for both link prediction + mask cell modelling, since it would give more insight about the dataset's performance on both link prediction and MCM. The model is trained by using these decoders to perform the desired prediction tasks to achieve results. This model is used to evaluate the metrics graph sparsity, number of missing cells, and modularity. The model is evaluated with 3 different metrics; MRR (Mean Reciprocal Rank), Accuracy, and RMSE (Root Squared Mean Error). MRR is a metric used to evaluate link prediction performance (1 being perfect link prediction), Accuracy shows how accurately the model performs Mask Cell Modeling (MCM) tasks, and RMSE indicates the error in the model's performance during MCM tasks, with lower values indicating better performance.

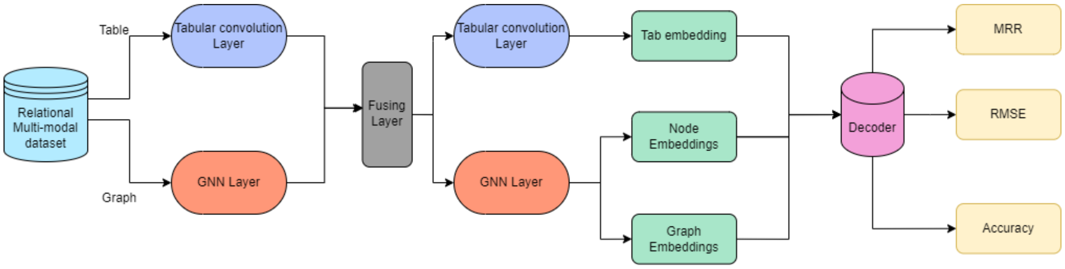

Figure 1: Representation of the first model

The second model which is used to evaluate the effect of length of the text on the model has a more basic high-level structure, where the RoBERTa [5] text embeddings are created at the start, and sent to an FTTransformer [6] to complete a regression task, which outputs a Mean Squared Error (MSE).

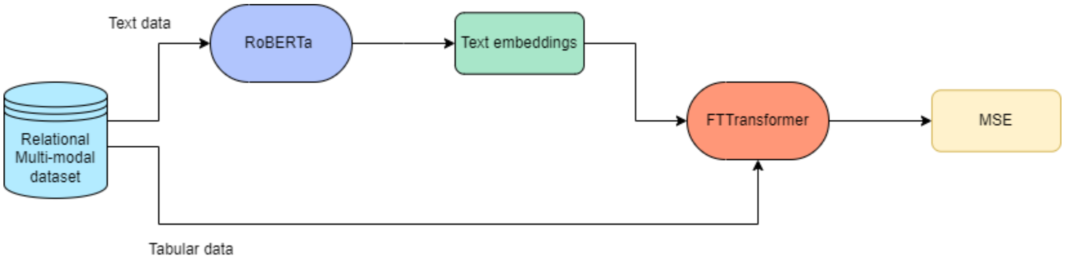

Figure 2: Representation of the second model

6.1 Graph Sparsity

Graph sparsity shows how many relationships there are, and how complex the graph is. Therefore, it can effect the model's performance directly, especially on link prediction related tasks.

To explore the impact of the density/sparsity of the graph, IBM Transactions dataset is used, since it provides a great structure to apply link prediction tasks, and also has a high sparsity rate on the full dataset. A random of 500 thousand rows are taken from the dataset, to make the computation more efficient.

A good way to make a graph denser is by adding more edges to it, however in our case, the graph nodes and edges has many features, so creating synthetic edges could cause more problems if it is not done properly. In this experiment, the approach is to increase the density by finding different denser subsets within the graph, in increasing density. However, it could also be important to keep the dataset size and label distribution similar, since it could also affect the performance of the model.

Sparsity	**MRR ↑**	**Accuracy ↑**	**RMSE ↓**
0.9999	**0.066**	**0.838**	0.110
0.8264	0.054	0.821	0.134
0.7452	0.056	0.792	0.148
0.6623	0.052	0.812	**0.166**

Table 6.1: Sparsity of the graph and the outputs of the model after 10 epochs

6.2 Number of missing cells

Missing data can occur for numerous reasons including data collection errors or privacy concerns. That's why it is important to learn about how it effects the model's performance to see to what extent the datasets with missing data is usable for machine learning tasks.

To explore this, the IBM Transactions dataset is used, since there are many columns that we can create missing cells, and does not contain text. So it can be used for even simple machine learning tasks. The dataset has columns like amount received (transaction amount) and receiving currency, which can be treated as missing values, and con be replaced with placeholder values. In this case, the categorical columns are replaced with "None" while numerical columns are replaced with "-1". The initial dataset doesn't have any missing cells, therefore this experiment will compare different rates of missing values.

Missing cells (Percentage)	**Number of Rows**	**MRR ↑**	**Accuracy ↑**	**RMSE ↓**
0%	500000	0.066	**0.838**	**0.110**
10%	500000	0.069	0.834	0.114
30%	500000	0.068	0.811	0.134
50%	500000	**0.071**	0.815	0.131

Table 6.2: Number of missing cells in the dataset and the outputs of the model after 10 epochs

6.3 Modularity

High modularity indicates a strong presence of clusters with dense connections internally and sparser connections between them. This can be crucial for understanding the quality of network partitions and can influence the performance of machine learning algorithms.

To investigate the impact of modularity on the graph, we use the IBM Transactions dataset, which is ideal for exploring various graph properties due to its complex structure. The dataset is divided into different clusters to study the effects of varying modularity levels. A random subset of 500 thousand rows is selected to maintain computational efficiency.

There are several different algorithms to calculate the modularity of a graph, such as the Louvain Method[3] and Spectral Optimization[15]. Hovewer, in this experiment, Label Propagation algorithm[17] is used, since it is faster to compute and can be easily done with a NetworkX function[7]. For changing the modularity, there are several ways to create a new subgraph like removing edges that connect nodes in different communities, and adding edges within or between communities. In this experiment, the first approach will be used, since in the second approach, adding synthetic edges to the graph might also effect it's quality, which can lead to different results. Using the first approach will help us evaluate the decoder's effectiveness in identifying community structures and understanding how modularity influences the performance on link prediction and MCM.

Modularity	**Number of rows**	**MRR ↑**	**Accuracy ↑**	**RMSE ↓**
0.5225	1500000	**0.3386**	**0.843**	**0,101**
0.6472	750000	0.066	0.838	0.110
0.7233	500000	0.061	0.813	0.116

Table 6.3: Text lengths of the data in the subsets and the MSE after 11 epochs

6.4 Text Length

Overall text length given might affect how the LLMs perform, which would change the outcome of the model. This experiment therefore focuses on dividing the Amazon-Fashion dataset into different subsets with different average text lengths properly. These subsets will then be used to train the model separately, which then they would give us insight about how the model performs.

The dataset is divided into 3 subsets with 100 thousand rows each, which includes reviews where the text length is smaller than 10, reviews where the text length is larger than 70, and reviews where the text length is in between.

The mean squared error (MSE) is calculated at the end of each epoch, which is used as a metric of performance of the model. Because of the computational and time constraints, only 11 epochs were taken into consideration as the final result of this experiment.

Text length	**Number of rows**	**MSE↓**
Any	100000	0.5523
<= 10	100000	0.3663
> 10 and < 70	100000	0.3934
>= 70	100000	**0.3502**

Table 6.4: Text lengths of the data in the subsets and the MSE after 11 epochs

7 Discussion

In this study, we examined the effects of graph sparsity, the number of missing cells, modularity, and text length on the performance of machine learning models. The information derived from these experiments are important for understanding the importance of these factors in the context of machine learning tasks, since they can guide us about creating new great datasets.

7.1 Graph Sparsity

Our results show a small relationship between graph sparsity and model performance. While extremely sparse graphs (sparsity 0.9999) resulted in the highest accuracy (0.838) and the lowest RMSE (0.110), this was not the case as sparsity decreased. For example, at a sparsity of 0.6623, the model's accuracy dropped to 0.812 and RMSE increased to 0.166.

One possible reason for these results could be that sparse graphs, despite having fewer connections, they may still capture relationships that are important for correct predictions. However, as we make the graph denser, the model could start focusing on redundant or less informative edges that do not contribute to the predictive power too much. This noise can obscure useful patterns and potentially lead to over-fitting, where the model performs good on the training data but bad on unseen data.

7.2 Number of Missing Cells

The presence of missing cells had a somewhat counter intuitive effect on model performance. Up to 50% missing cells, the model had a high accuracy (0.815) and low RMSE (0.131). Interestingly, there was a slight improvement in MRR as the percentage of missing cells increased.

This might be because the model could rely more on the patterns. When the model encounters missing data, it might rely more on the patterns within the data rather than fitting to potentially correlations present in full datasets. Additionally, the imputation techniques used to handle missing data could play an important role. Effective imputation methods can reduce the negative impact of missing values, keeping the integrity of the dataset and supporting strong model performance. If the experiment is repeated again using different imputation techniques, the performance could be worse, or better. This is why it could be a great idea to repeat the experiment with different techniques, to see which technique works the best and the worst.

7.3 Modularity

The new data on modularity presents an interesting scenario. Higher modularity levels (0.6472) resulted in high accuracy (0.838) but also a slight increase in RMSE (0.110) compared to lower modularity (0.5225), which showed an accuracy of 0.843 and a lower RMSE (0.101). The highest modularity level (0.7233) led to a decrease in accuracy (0.813) and an increase in RMSE (0.116).

This suggests that while modularity can improve the learning of meaningful patterns by creating defined clusters, there is a trade-off. Higher modularity might create complexity by creating more defined clusters that the model finds difficult to understand. The increased complexity can result in higher variability in model predictions, therefore can increase RMSE. Additionally, it is also seen that the MRR decreased with increasing modularity. This could be because of the size of the graph instead of modularity. In smaller datasets, it is likely that the neighbourhoods are much smaller when sampling edges for link prediction, which could result in poor performance.

7.4 Text Length

Text length showed a clear influence on model performance, with longer texts (>= 70 words) yielding the lowest MSE (0.3502). In contrast, shorter texts (<= 10 words) resulted in a higher MSE (0.3663). This shows that richer textual information provides better context, enabling the model to make more accurate predictions.

The overall higher MSE for text-based tasks compared to graph-based tasks shows the potential limitations in the model's ability to effectively utilize textual data. This could be due to the inherent change and noise in text data, which can create challenges for pattern extraction. Longer texts might provide more context, but they also could introduce more irrelevant or noisy data, which the model needs to learn to filter out effectively. RoBERTa embeddings and FTTransformer appears to handle this reasonably well, but improvements could be explored in future work.

8 Conclusion and Future Improvements

The goal of this research was to explore more about what makes a great dataset. It focused on extracting some features that could influence the quality of datasets used in machine learning, and find out if we can describe a dataset's usefulness using those features. The features used mainly in the research were listed as graph sparsity, number of missing cells, modularity, and text length. The experiments conducted to provide insights into how these factors impact model performance and offer a foundation for future dataset construction and evaluation.

The results show that:

- **Graph Sparsity:** Extremely sparse graphs can keep in some important information, but increasing density does not directly improve performance due to potential noise and complexity. The introduction of redundant or less informative edges in denser graphs may obscure useful patterns and lead to over-fitting.
- **Number of Missing Cells:** The model shows resilience to missing data, maintaining high accuracy and low RMSE even with up to 50% missing cells. The model's ability to generalize from incomplete data, or the imputation techniques could contribute to this resilience.
- **Modularity:** Higher modularity generally improves model performance by facilitating meaningful pattern learning, however the methods of changing modularity can introduce challenges. The balance between benefiting from clear community structures and managing the complexity they introduce is crucial.
- **Text Length:** Longer text improves the model performance by providing more contextual information, but the overall noise in text data can create challenges. The model needs to effectively filter out irrelevant information to leverage the benefits of longer texts.

These findings show the importance of carefully considering graph and data characteristics when developing machine learning models for complex predictive tasks. By understanding and optimizing these factors, we can improve the accuracy of models, making them better suited for real-world applications, and even create better datasets.

Building on these insights, future research could explore other techniques for handling graph sparsity and missing data, such as trying out new data imputation methods and graph augmentation strategies. Additionally, further investigation into the relationship between text length and other linguistic features could result in improvements in text-based model performance. Optimizing modularity through more refined clustering algorithms and exploring their impact on various graph-based task could also present a great approach to improve the model capability. Apart from these, for each metric, multiple experiments could be done, with using different datasets and different variables, to learn more about how the metrics actually influence the model. And finally, other graph/dataset metrics like diameter of a graph and degree centrality could be evaluated as well, to determine what makes a dataset great.

Overall, this study lays a foundation for future research about the graph and text characteristics in machine learning, guiding the development of more effective and resilient models.

9 Responsible Research

Conducting responsible research is an important step that involves ensuring ethical considerations and reproducibility that could affect the outcomes of the study. This section outlines the measures taken to uphold these principles.

9.1 Concerns

While and after conducting the research, there are several things that we need to be aware of, to reduce the possible negative effects of this research.

9.1.1 Energy Consumption

The research requires too much amount computational resources and therefore energy consumption, for training complex machine learning models. We acknowledge this environmental impact and therefore didn't use any unnecessary computational power, and tried to do everything locally as much as possible.

9.1.2 Bias

Bias in machine learning can arise from various sources, including the data, the models, and the experimental design. To mitigate bias, the following steps were taken:

- **Balanced Sampling:** In experiments involving subsets of data, care was taken to ensure that the samples were balanced and representative of the overall dataset depending on the task.
- **Algorithmic Fairness:** The models were evaluated using multiple metrics (MRR, Accuracy, RMSE, MSE) to ensure a fair assessment of their performance. This multifaceted evaluation helps in identifying any biases that might arise from relying on a single metric.

9.2 Reproducibility

Ensuring the reproducibility of our experiments is important to validate the results and enable other researchers to build upon our work. We have taken several steps to facilitate reproducibility:

9.2.1 Data Availability

All datasets used in this study are publicly available and can be accessed from the following sources:

- **OGB Arxiv (Open Graph Benchmark):** `https://ogb.stanford.edu/docs/nodeprop/#ogbn-arxiv`
- **Amazon-Fashion:** `https://jmcauley.ucsd.edu/data/amazon/`
- **IBM Transactions for Anti Money Laundering:** `https://www.kaggle.com/datasets/ealtman2019/ibm-transactions-for-anti-money-laundering-aml`
- **Ethereum Phishing Transaction Network:** `https://www.kaggle.com/datasets/xblock/ethereum-phishing-transaction-network`

9.2.2 Code and Implementation

The code used to preprocess the data, train the models, and a README file to help you setup and reproduce the results is available in the repository, which is also provided with required comments and documentation.

References

[1] Imran Razzak Adnan Ejaz and Saeed Anwar. A survey on dataset quality in machine learning. *Engineering Applications of Artificial Intelligence*, 2023.

[2] Eric Altman. Ibm transactions for anti money laundering. `https://www.kaggle.com/datasets/ealtman2019/ibm-transactions-for-anti-money-laundering-aml/data`, 2024.

[3] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[5] Hugging Face. Roberta model documentation, 2024. Accessed: 2024-06-17.

[6] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34:18932–18943, 2021.

[7] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Networkx, 2008. Python library for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks.

[8] Wei Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hengrui Luo, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *arXiv preprint arXiv:2005.00687*, 2020.

[9] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *arXiv preprint arXiv:2005.00687*, 2020.

[10] Julian McAuley Jianmo Ni, Jiacheng Li. Justifying recommendations using distantly-labeled reviews and fined-grained aspects. 2019.

[11] Yann LeCun, LÃ©on Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86, pages 2278–2324, 1998.

[12] Jure Leskovec and Andrej Krevl. Snap datasets: Stanford large network dataset collection. Technical report, Stanford University, 2014.

[13] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[14] Author Names. The effects of data quality on machine learning algorithms. *MITIQ*, Year. Accessed: 2024-06-07.

[15] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

[16] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, pages 8024–8035, 2019.

[17] Usha Nandini Raghavan, RÃ©ka Albert, and Soundar Kumara. Near linear time algorithm to detect community structures in large-scale networks. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pages 1061–1066. SIAM, 2007.

[18] xblock. Ethereum phishing transaction network. `https://www.kaggle.com/datasets/xblock/ethereum-phishing-transaction-network`, 2021.