Analyzing the Social Implications of Outlier Removal on Predictive Models

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Anh Nguyen

Spring 2023

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Pedro A. P. Francisco, Department of Engineering and Society

Introduction

Machine learning has become a popular computer science field in recent years, and many everyday applications implement machine learning models for a variety of purposes. Recently-developed predictive models incorporate machine learning algorithms and models since these have thorough data pre-processing and data processing steps which lead to the model making accurate predictions. As more research in the machine learning field continues, more focus and emphasis is put on the data collection, the model training, and the results of these predictive models. These new focuses draw attention away from how the data is cleaned, especially when studies implement outlier removal. There is a lack of transparency when outlier removal is applied, which raises the question of who and what is being excluded from these models and if these models are accurately representing society? It is essential to study outlier removal during data cleaning in order to better understand bias in machine learning models and their predictions, since the bias comes from the data the model gets trained on. By removing outliers, we run into the possibility of excluding different social groups, so this study will analyze the social implications of outlier removal within different types of predictive models with machine learning models and algorithms in a variety of fields. This study will also use the theoretical framework ethics of care and the relational view to further understand the role of outliers in different sets of data.

Background and significance: Why should we care about outlier removal?

Outliers are prevalent and are taught throughout years of education. Among the majority of researchers and students today, outlier removal is done regularly during data pre-processing or data cleaning. Outlier removal is especially prevalent in computer science classes and fields like machine learning and artificial intelligence (Caton et al., 2022). In these fields and classes, there

is always a lot of study and debate on outliers' role in the predictive model. Most of the times, data is just seen as numbers and outliers in the data are just dropped during data pre-processing without careful identification. The outliers are dropped only after looking at how far the data point does not fit in with the average and usually no other context or characteristics. Sometimes, when the data is grouped, outliers of the groups are dropped even though the outliers are not considered outliers when looking at the data as a bigger picture. Many disagree with leaving outliers in data since outliers can skew the results of a predictive model and also lead to inaccurate predictions (Liu et al., 2021). In order to better understand the researchers' point of view on outliers and outliers in datasets, we need to consider what the outliers represent within the dataset.

When managing outliers, it is common to see studies attach a negative connotation to outliers in data, which encourages and enables outlier removal in datasets. In a past study done by Bollen (1988), he looked at outlier detection and if the management of outliers affected data negatively. In another study by Liu et al. (2021), they also refer to outliers as having a negative impact on data. These studies raise the question of what makes an outlier have a negative impact on data? In 2019, Osborne and Overbay did a study to analyze and summarize the different types of outliers in data. Outliers in data can come in many different forms and they can have different meanings in data (Osborne and Overbay, 2019). Osborne and Overbay identified some outliers as data errors, sampling errors, standardization errors, sampling assumption errors, and more. Most of these outliers can be grouped as errors during the data collection phase, like collecting data from a person who does not meet the requirements of the study (Zijlstra et al., 2011). Knowing these types of outliers and identifying them as outliers during data collection is useful when analyzing other studies. If the outliers are errors from data collection, researchers'

argument to remove these kinds of outliers makes sense since these outliers do not represent the population the model is supposed to be made for and would impact the data negatively in these cases. Osborne and Overbay (2019) also mention that generally outliers are "…a data point that is far outside the norm for a variable or populations" (p. 1). However, if outliers are just generalized as so, and are removed, then the results produced by the models using this data will not be accurately representing the situation or population.

Research methods: literature review and theoretical frameworks

To understand the role of outliers in datasets and predictive models, this study looks into various fields using different types of data. By analyzing different models in different fields, we can see how data is perceived by different researchers and observe different motivations behind outlier removal in different types of data. Not all of these models will be predictive models, but the models in these studies still rely on correlating and training on a dataset. We can also see how researchers portray the limitations of their model after outlier removal and what future research is needed on their models. How do they make up for the data that gets cut out from their models, and do they claim that their models do not exclude different groups in society? By analyzing different fields, we can also see how different fields treat outlier removal and data cleaning. Are researchers transparent with the data they use and the way they clean the data? These questions and observations will be reviewed in studies using medical data and financial data. The models using medical data are trained on data from patients, and the models using financial data are trained on data that are not directly linked to a person. By analyzing these different types of data, we can observe how the researchers approach cleaning different types of data and see how their fields may view the data.

The theoretical frameworks applied after the literature review are ethics of care and the relational view. After the literature review, these theoretical frameworks will be applied to show how to help with the exclusion or other social implications the models may have. These two frameworks will also help with addressing the limitations and exclusions in the models from their outlier removal step.

The ethics of care theoretical framework was chosen and works well for this study because, at least in the context of medical and patient data, it will help us better understand the data by making us look at who is included in the models. Taylor (2020) conducted a study on the Covid-19 pandemic and policymakers. Taylor stated that policymakers need to understand why different groups in society could not follow guidelines implemented by the government and how policy makers should approach the situation with ethics of care. By seeing different groups in society as different types of people, not just the general population, and understanding why they cannot comply with the guidelines, the policy makers can change the guidelines for them to better avoid spreading Covid-19. We can use ethics of care in this study to look at the data as people and not as numbers or data. This will force us to look at the data as people. How does removing this person from the data change the model and does that end up making the model exclude a group of people in society? This will help us see the social implications of outlier removal in the different studies and also propose a different way of perceiving data in the medical data for future studies.

The relational view theoretical framework also works well for this study, because it requires us to understand the context of outliers in the data and in different situations. According to Lionelli (2019), "…data are 'relational': in other words, the objects that best serve as data can change depending on the standards, goals and methods used to generate, process and interpret

those objects as evidence" (pg. 8). The context of the outliers in the data can range from sampling errors or other errors during data collection to outliers that just do not correlate and fit the average of the rest of the data. Using this framework can also help us see the limitations of models that use outlier removal to get more accurate predictions or results. Limitations of models will require other experts in different fields or future research to address cases that exclude outliers, which is another reason why this theoretical framework is useful. This theoretical framework can also explain the differences between different fields perceiving and treating data during data cleaning. For example, the relational view will help with understanding the circumstances of the medical data compared to the circumstances of the financial data and how outlier removal when dealing with these two types of data mean something different from one another.

Results and discussion

The first study we will reference is an outlier removal study done by Uzun et al. (2022) exploring outliers and their affect on predictive models for medical disease diagnosis. In this study, the authors used five different machine learning models to predict diseases in patients. Here, we will have a chance to see how outlier removal affects various models with different machine learning algorithms. The authors also used four different datasets with different types of disease data and characteristics. One of the datasets was made up of digitalized images and the other datasets were made up of different numerical or categorical instances. Each of these datasets also had different diseases and different populations, but all the medical data was patient data. The authors explained the datasets and the way they pre-processed the data.

The first observation Uzun et al. (2022) made for these datasets was whether the datasets had null or missing values in the entries. If this was the case, they removed those entries. Entries

with missing values can be seen as outliers, and removing them can be beneficial to the overall study or model since missing values is incomplete data. In this case, it seemed fine for the authors to remove these patient's data since it did not represent the patient properly and these can be seen as errors that happened during data collection. For other outliers, they first identified them using boxplots that summarized the whole dataset, and then they removed the outliers or points that went beyond the scope of the average of the data. At the beginning of their paper they identified outliers as contextual outliers, collective outliers, or point outliers. Contextual outliers are datapoints that are outliers in a certain context. Collective outliers are a group of datapoints that are outliers. Point outliers are individual data points that are outliers. Although they mentioned that this was their outlier detection method, in the data pre-processing step, they did not clarify which identification of outliers they were removing. They just stated that outliers were removed using the boxplot which can be confusing for the reader since there are different types of outliers that should be removed in different cases, but the authors just mentioned generally that they removed the outliers.

When testing all the different models, the authors mention that outlier removal had little influence on the accuracy of their machine learning models when predicting and diagnosing the diseases. They also mention that outlier removal, while seen as a necessary step in the machine learning field, may not be necessary when it comes to medical data. Some of the entries may seem out of range when grouping the data, but in reality, those extreme cases represent an absence or presence of the disease. These cases can represent a different threshold of patient data that do not follow the norm.

In this study, this is a case where ethics of care was not used. Throughout the study, the data is looked at as just entries, numerical values, and categorical values. The data is not seen as

the patient themselves. The conclusion the authors came to still focused on the outliers representing a larger presence or an absence of the disease, but the authors do not mention say anything about how we should keep the complete outlier entries in the data because the data is a person and not including them implies excluding a whole ancestry of different people and their medical data. The authors also put more emphasis on how the outlier removal does not affect the performance of the model. If outlier removal had a large impact, it is likely that their conclusion would be different from outlier removal possibly being unnecessary. Relational view can be seen here as the authors mention what the outliers mean in the medical field and how handling medical data is different from handling other data. They explain the circumstance of outliers in medical data and also explain medical data generally as having more implications than just being data points.

The second outlier removal study I looked at was a past study done by Pollet and van der Meij (2017) exploring outlier removal in hormonal data and the impact outlier removal has on significance testing in testosterone data. This study does not reference predictive models, but the authors provide insight on hormone research and express a need for better outlier management in this field. Their study concluded that in hormonal research, removing outliers from the data makes a big difference on the statistics of the data compared to keeping the outliers in the data. At the end of their paper, they also address potential solutions to future researches on how to be transparent with the outliers they are removing. They also encourage researchers to make separate models for removing or keeping outliers and to share the results of both models. This study is an example where ethics of care and the relational view are being implemented. The authors explain how with each hormone research dataset, the outliers need to be carefully considered since the data represents people. Removing outliers in order to get a higher

performance or more consistent statistics cuts out the people from the dataset which can make the research biased against the people excluded. They also used the relational view to show that you can have a model with outliers, but you just need to explain the meaning behind the results. We can use the relational view to draw relation between the outliers and the hormone research field and keep the outliers in the model or the statistics.

Both the study done by Uzun et al. (2022) and the study done by Pollet and van der Meij (2017) use patient data in the medical field, but the way they came about their conclusions is quite different. Uzun et al. brought insight on managing outliers in the medical field for disease diagnosis and how outlier removal may not be as necessary in the medical field as the machine learning field emphasizes. However, Uzun et al. did not put much emphasize on looking at what the data represents like Pollet and van der Meij did. Pollet and van der Meij focused on why to remove and when to remove. They also mentioned how researchers should take extra care to manage outliers instead of removing them altogether for better performing models. They also mention how researchers should be clear and transparent about their data and on their data cleaning process, which both papers had some ambiguity in. Models where outliers are not carefully managed result in a misleading assessment of patients which can also lead to other complications in hospitals, doctor offices, and the overall treatment process. Overall, both studies still push against the usual outlier removal necessity that does not consider the context of the data.

The last study I wanted to reference is not in the medical field. Unlike the studies mentioned earlier, this one contains data that does not represent people. This study is more focused in modeling financial data. The study by Dutta (2018) examines the modeling of the carbon emission market. Dutta (2018) states that “the detection of outliers in financial time series

is important, since the presence of such extreme observations can bias the estimation of parameters and also lead to poor forecasts and invalid inferences" (p. 2779). Similar to the other two studies, Dutta puts emphasis on the impact on the performance of the model and not the context of the data. In the carbon emission market model with outlier removal, Dutta observes that its analysis of the emission market is not accurate. This is because of the way the data was pre-processed. Unlike the previous studies, the outlier removal process was much more unclear. The outliers removed seemed to be based on numerical values or the numerical data points. The spikes in the carbon emission market were removed in order for the model to provide better predictions using the standardized data. This is not an accurate representation of reality as there are spikes in the emission market. Since this model cannot account for these spikes or cannot provide other ways to use the extreme cases, it would need economists, policymakers, and etc. to account for the outliers. Dutta does mention this limitation of the model that removed the outliers. If outliers are not managed properly in this case, it can result in misleading analysis on the carbon emission market and implementation of polices that do not properly mitigate the risks represented in the emission market.

Dutta's study does not really use ethics of care or the relational view theoretical frameworks. The only time relational view can kind of be referenced is when Dutta explains the limitations of the model and how experts in other fields would be needed to consider the unaccounted outliers. In this case, ethics of care would not be applied to the data and outliers but the circumstances of the carbon emission market and who the market affects. We would look at the policies that surround the carbon emission market and why policies may have to change in different times or for different people. For the relational view, we would want to take into account other social affairs happening in the world and other economies. We would also want to

reference the time of the spikes and what other factors could be contributing to the spikes and outliers in the emission market. This was one of the studies that do not use data representing people, but still requires context on the data and other factors in society that could relate to the study.

Conclusion

After referencing different studies from various fields that train predictive or normal models on datasets, data cleaning and outlier removal need to be approached in a different manner than what the field currently emphasizes. In many fields requiring data, like machine learning, outlier removal is a no brainer. Outliers are detected if they are outside some specified range and removed without considering their context. If there is a specific identification or label for the outliers, the reasons for their removal are usually ambiguous and authors will favor removing them for the overall performance of their research. Without approaching the outliers with ethics of care or the relational view theoretical frameworks, it is easy to disregard outliers. Disregarding outliers can lead to models not accurately representing data, models excluding different groups in society, and models unable to provide accurate results for the situations they are made for. Should outliers no longer be removed in data? Not necessarily. It makes sense to remove outliers that are errors during data collection. Is there a specific way to manage outliers so that predictive models produce accurate results? This will vary with each model, but there are extra steps researchers can take to combat this. They can keep outliers in their models and explain the accuracy of their models based on the outliers. However they choose to identify and incorporate the outliers into their model, or if they choose to remove the outliers, they need to be transparent about what they do. The current lack of transparency when dealing with outliers and cleaning data needs to change. While the advances in computer science fields and models can be

extremely useful, extra steps should be taken to make sure that these models can be applicable in the real world. Cutting corners during data pre-processing and making these models for the sake of better performance will come back to cause more harm than the help the model should be providing.

References

Bollen, K. A. (1988). “If you ignore outliers, will they go away? ”: A response to gasiorowski. *Comparative Political Studies*, 20(4), 516–522. https://doi.org/10.1177/0010414088020004005

Caton, S., Malisetty, S., & Haas, C. (2022). Impact of Imputation Strategies on Fairness in Machine Learning. *Journal of Artificial Intelligence Research*, *74*. https://doi.org/10.1613/jair.1.13197

Dutta, A. (2018). Modeling and forecasting the volatility of carbon emission market: The role of outliers, time-varying jumps and oil price risk. *Journal of Cleaner Production*, 172, 2773–2781. https://doi.org/10.1016/j.jclepro.2017.11.135

Leonelli, S. (2019). The challenges of big data biology. *ELife*, *8*, e47381. https://doi.org/10.7554/eLife.47381

Liu, H., Li, J., Wu, Y., & Fu, Y. (2021). Clustering With Outlier Removal. *IEEE Transactions on Knowledge and Data Engineering*, *33*(6), 2369–2379. https://doi.org/10.1109/TKDE.2019.2954317

Osborne, J., & Overbay, A. (2019). The power of outliers (And why researchers should always check for them). *Practical Assessment, Research, and Evaluation*, 9(1). https://doi.org/https://doi.org/10.7275/qf69-7k43

Pollet, T. V., & van der Meij, L. (2017). To remove or not to remove: The impact of outlier handling on significance testing in testosterone data. *Adaptive Human Behavior and Physiology*, 3(1), 43–60. https://doi.org/10.1007/s40750-016-0050-z

Taylor, L. (2020). The price of certainty: How the politics of pandemic data demand an ethics of care. *Big Data & Society*, *7*(2), 2053951720942539. https://doi.org/10.1177/2053951720942539

Uzun Ozsahin, D., Taiwo Mustapha, M., Saleh Mubarak, A., Said Ameen, Z., & Uzun, B. (2022). Impact of outliers and dimensionality reduction on the performance of predictive models for medical disease diagnosis. *2022 International Conference on Artificial Intelligence in Everything (AIE)*, 79–86. https://doi.org/10.1109/AIE57029.2022.00023

Zijlstra, W. P., van der Ark, L. A., & Sijtsma, K. (2011). Outliers in Questionnaire Data: Can They Be Detected and Should They Be Removed? *Journal of Educational and Behavioral Statistics*, *36*(2), 186–212. https://doi.org/10.3102/1076998610366263