Using Machine Learning K-Means Clustering to Comprehend California Housing Prices

Analyzing the Social Implications of Outlier Removal on Predictive Models

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Anh Nguyen

November 4, 2022

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.

ADVISORS

Caitlin Wylie, Department of Engineering and Society

Briana Morrison, Department of Computer Science

Introduction

Machine learning has become a popular computer science field in recent years, and many applications used today implement machine learning algorithms for a variety of purposes. One type of machine learning algorithms is clustering which can be applied to many different datasets to better understand trends in society and predict fairly accurate outcomes. For my technical topic, I plan to develop a program using the machine learning K-means clustering algorithm to group California housing regions. This program can help people interested in living in California better understand California housing prices and trends, which can also lend a hand in better understanding the California housing market. I will also be researching what factors affect the K-means algorithm performance in order to better understand the algorithm and see it's compatibility with grouping housing data.

Predictive models based on machine learning algorithms rely heavily on correlating characteristics in datasets. Cleaning these datasets for the models to be trained on is a crucial step for these models; and for my STS topic, I plan to analyze the removal of outliers from datasets during this step and the social implications it has on these predictive models. Outliers in data pose problems because they can skew predictive models, but outlier removal raises the social and ethical question of whether the data can still accurately represent the situation or population after the outliers are removed (Bollen, 1988). I plan to analyze ways previous studies have dealt with different types of outliers, what social groups can be affected from outlier removal in those studies or situations, and whether outlier removal misrepresents populations or the data.

Technical Topic

Whether it be the weather or the high density of technology companies, there are many factors that attract people to living in California (Rees, 2022). However, the downside of living in California is the exorbitant housing prices. Housing prices have increased a lot due to the high demand and limited supply in housing during Covid-19 (Riquier and Witkowski, 2022). This makes the clustering housing program a useful tool for people interested in buying a home or investing in property, especially in places like California where the housing prices are already high. It would give insight into what characteristics affect housing prices in California.

The program implementation can be broken up into four parts: loading the dataset, pre-processing or cleaning the data, implementing the K-means algorithm, and clustering the housing into regions. The K-means clustering algorithm is an unsupervised machine learning algorithm that groups data points into a number of K clusters based on a distance metric from the data point to the center of the cluster. I plan on programming the algorithm in Python and using various Python packages to clean the California housing dataset from Kaggle, a well-known and popular platform where many data scientists publish or find datasets, and implement the K-means algorithm. I will also look into what components affect the K-means algorithm performance. The first component is the distance metric used to calculate the centroids of the clusters. Gupta et al. (2022) argue that the Minkowski distance is a better metric for the K-means algorithm compared to Euclidean distance metric because the Minkowski distance metric can be seen as a generalized version of the Euclidenan distance metric. The second component I plan to research is the number of clusters that works best with grouping the data. I will start with clustering the data in two groups and going up to ten groups. Looking at these components will help give insight on how the K-means algorithm works and whether a large number of clusters is better than a small number of clusters.

Part of my technical research includes looking into previous works related to grouping housing and dealing with housing data. One previous study done by Truong et al. (2022) is about how to use machine learning techniques to predict housing prices that don't just consider past sale prices. The authors argue that only predicting house prices using the housing price index (HPI) isn't accurate and that other features need to be considered when predicting house prices. Price and all the other characteristics in the dataset are relvant in grouping, so having an idea of some characterstic groupings is helpful for my research. The study by Truong et al. (2022) is similar to the study by Gupta et al. (2022) because both use machine learning to group characteristics of data. These studies in peer reviewed journals will help with giving insights on how to clean data for the K-means algorithm and how to implement the algorithm to result in a better performance.

STS Topic

My STS topic takes a step back from the machine learning algorithm implementation, and focuses on the step before that: data cleaning and processing. In this step specifically, for my STS topic, I will analyze the impact of outlier removal on predictive and machine learning models. In the machine learning field, there is ongoing debate about how outliers are dealt with in the data cleaning process of programs (Osborne and Overbay, 2019). Osborne and Overbay argue that in order to know what to do with outliers, you must understand the context of outliers in the dataset before managing outliers. Osborne and Overbay explain how there are many types of outliers. For example, some outliers are intentional and some are from sampling error. Understanding outliers within data can lead to seeing the social implications of removing them. What groups of people are not accounted for in the models, and can the models be said to provide accurate predictions if they don't represent the population they're made for? This source is in a peer-reviewed journal and is useful for my STS research since it provides context on types of outliers and discusses methods on dealing with outliers. The authors also compare the performance of removing outliers and keeping outliers in various studies using different datasets.

In order to understand the effect of outlier removal on predictive models using machine learning, I will look at peer reviewed journals with studies that research the role of outliers in modeling financial data. One specific study of relevance examines the modeling of the carbon emission market done by Anupam Dutta (2018). 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). Dutta mentions this, but finds ways to manage outliers in the model. The model detects and accounts for outliers with an outlier-correction feature. This model can help with developing risk management when it comes to making predictions for the carbon emission market. Dutta does study a model with outlier removal and mentions that its analysis of the emission market is not accurate. The model would need economsits, policymakers, and etc. to account for 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 don’t properly mitigate the risks represented in the emission market.

Another past study I will use in my research looks at is managing outlier removal and its affect in the medical field. Pollet and van der Meij (2017) explored outlier removal in hormonal data and the impact outlier removal has on the conclusions. The authors argue that outliers in hormonal data need to be managed carefully and they propose solutions for managing outliers (Pollet and van der Meij, 2017). Some of the social impacts that could come with an inaccurate model that didn’t manage the outlier data properly is misleading assessment of endocrinology and endocrine or hormone disorders for patients which can lead to other complications in hospitals, doctor offices, and the treatment process. Another source I plan to look at is outliers and their affect on predictive models when it comes to medical disease diagnosis (Uzun et al., 2022). The authors use five different machine learning models, so this will help with better understanding how outliers influence each model. The authors argue that in the context of medical data, outliers should not be removed. They present that keeping the outliers in their model did not change the results of medical disease diagnosis. Since these outliers are referenced in the model, they can give broad and generalized predictions that warn patients of potential medical diagnoses in the future. This is better than ignoring the outlier altogether and missing the potential early diagnosis for patients. I agree with the authors on the importance of outliers in medical data because I think getting a false positive result allows room to observe the diagnosis and prevent complications in the treatment process.

I plan to draw on the ethics of care theoretical framework to show the relationships of outliers and how they help with understanding the predictive models or situations. Ethics of care works best for this analysis since understanding outliers in the context of the data can help build predictive models that do not marginalize outliers. Drawing on ethics of care can also help explain why certain outliers are dropped fom datasets and the context of the outliers in the models. I can apply this to the previous sources and Hickman et al.'s (2020) study about outlier removal using population data. This source addresses the exclusion or outlier problem in studies on healthy or unhealthy populations. Ethics of care can also help with analyzing this source and whether outlier removal from datasets misrepresent populations or groups within populations.

Conclusion

With machine learning algorithms being implemented in predictive models and commonly used applications, studying to improve them and studying which situations each algorithm works best in is relevant in today’s society. These algorithms are heavily reliant on correlating data, so researching the ethical side of outlier removal during the data processing step is also important. Using ethics of care, we can further understand the role of outliers in data and models. We can also use ethics of care to understand outlier groups in data and the context of those groups in society or in the field of study. Managing the data before applying the algorithm is relevant today since we want to make sure the models we make are ethical and do not marginalize certain groups in society. We want to make sure that the managed outliers accurately represent populations and the situations the predictive models are made for.

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

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

Gupta, M. K., & Chandra, P. (2021). Effects of similarity/distance metrics on k-means algorithm with respect to its applications in IoT and multimedia: A review. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-021-11255-7

Hickman, P. E., Koerbin, G., Potter, J. M., Glasgow, N., Cavanaugh, J. A., Abhayaratna, W. P., West, N. P., & Glasziou, P. (2020). Choice of statistical tools for outlier removal causes substantial changes in analyte reference intervals in healthy populations. *Clinical Chemistry*, 66(12), 1558–1561. https://doi.org/10.1093/clinchem/hvaa208

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

Rees, A. (2022). Top 10 best cities for tech jobs. *Bloom Institute of Technology*. https://www.bloomtech.com/article/top-10-best-cities-for-tech-jobs#:~:text=San%20Francisco%20has%20the%20largest,remains%20a%20top%20innovation%20hub.

Riquier, A., & Witkowski, R. (2022). Why are houses so expensive? *Forbes Advisor*. https://www.forbes.com/advisor/mortgages/real-estate/why-houses-are-expensive/

Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing price prediction via improved machine learning techniques. *Procedia Computer Science*, 174, 433–442. https://doi.org/10.1016/j.procs.2020.06.111

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