Thesis Project Portfolio

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

(Technical Report)

Analyzing the Social Implications of Outlier Removal on Predictive Models

(STS Research Paper)

An Undergraduate Thesis

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

University of Virginia • Charlottesville, Virginia

In Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Anh Nguyen

Spring, 2023

Department of Computer Science

Table of Contents

Sociotechnical Synthesis

California is known for its high housing prices and its high-demand for housing, and my capstone research addresses how people can better understand housing price trends in California. To better understand the housing price trends in California, I made a program that uses the K-means machine learning algorithm to cluster different characteristics in the California Housing dataset from Kaggle. This program gives insight on what houses in different regions of California have in common and how these factors can affect the prices of the houses. This program also explores the K-means clustering algorithm.

It is important to consider the human and social dimensions of my program. While my program can help give insight into characteristics that determine housing prices in California, it generalizes a lot of the characteristics and focuses more on clustering the houses into different regions. When looking at these limitations, we can see how my program would not be a technology that should be readily available to the public since it is lacking in different aspects. It also does not reference actual social cases where pricing is unequal and only looks at the data as numbers. To further analyze my capstone research, ethics of care and the relational view theoretical frameworks can be applied. Actor network theory would also work for analyzing since it can address human and non-human factors associated with housing prices and with the technology itself.

For my STS research, I am analyzing the social implications of outlier removal on predictive models. For my methodology, I conducted literature reviews and used ethics of care and the relational view theoretical frameworks to further analyze the different studies. Through my STS research, I explored how predictive models clean, pre-process, and process data. I also saw how outlier removal, specifically during pre-processing, is managed and the role of outliers

or how they are viewed in this field. When looking at both topics of research, we will be able to see how these predictive models end up with skewed or biased results based on how the models' training data is cleaned. We can also see what social groups end up being excluded from these models, the researchers' view on data in general, and whether or not the way this data is viewed needs to change.