BU

Department of Computer Science

Nathan Canterbury
Connor Gilheany
Dennis Nicolas

CS506 Data Science: Health Connector Report Analysis

Introduction

Our project partner, DiscoverASG, has been running a health insurance advertising and awareness campaign since 2016, specifically targeting Latino communities in Massachusetts, in an attempt to help close the ethnic gaps in health insurance coverage and prevent cyclic medical debt issues in minority communities. The advertising campaign placed window signs, broadcasted TV ads, and played radio ads in a number of cities in Massachusetts, many of them with dense Latino communities. DiscoverASG asked us to use the data science methods we've learned in order help determine if the advertising campaign was successful in reaching its target communities.

Data Reflection

The primary data sets used in our analysis were the Massachusetts Health Insurance Survey (MHIS), yearly surveys that collects information about health insurance coverage status, opinions on health insurance, income statistics, medical history, and more. The information collected in the MHIS is broken down by region, age, gender, and race. Another dataset we used was the Health Connector April Report provided by DiscoverASG, which contained the cumulative advertising data from DiscoverASG's aforementioned advertising campaign.

Two issues existed with the data:

1. The relationships between many of the data-points were disjoint, meaning many pockets of "comparable things" had a limited feature space. Working on a reasonably-sized dataspace required using a limited set of feature dimensions. Because of the socially-cognizant nature of the task, we settled on using regional feature spaces, coupled with income.

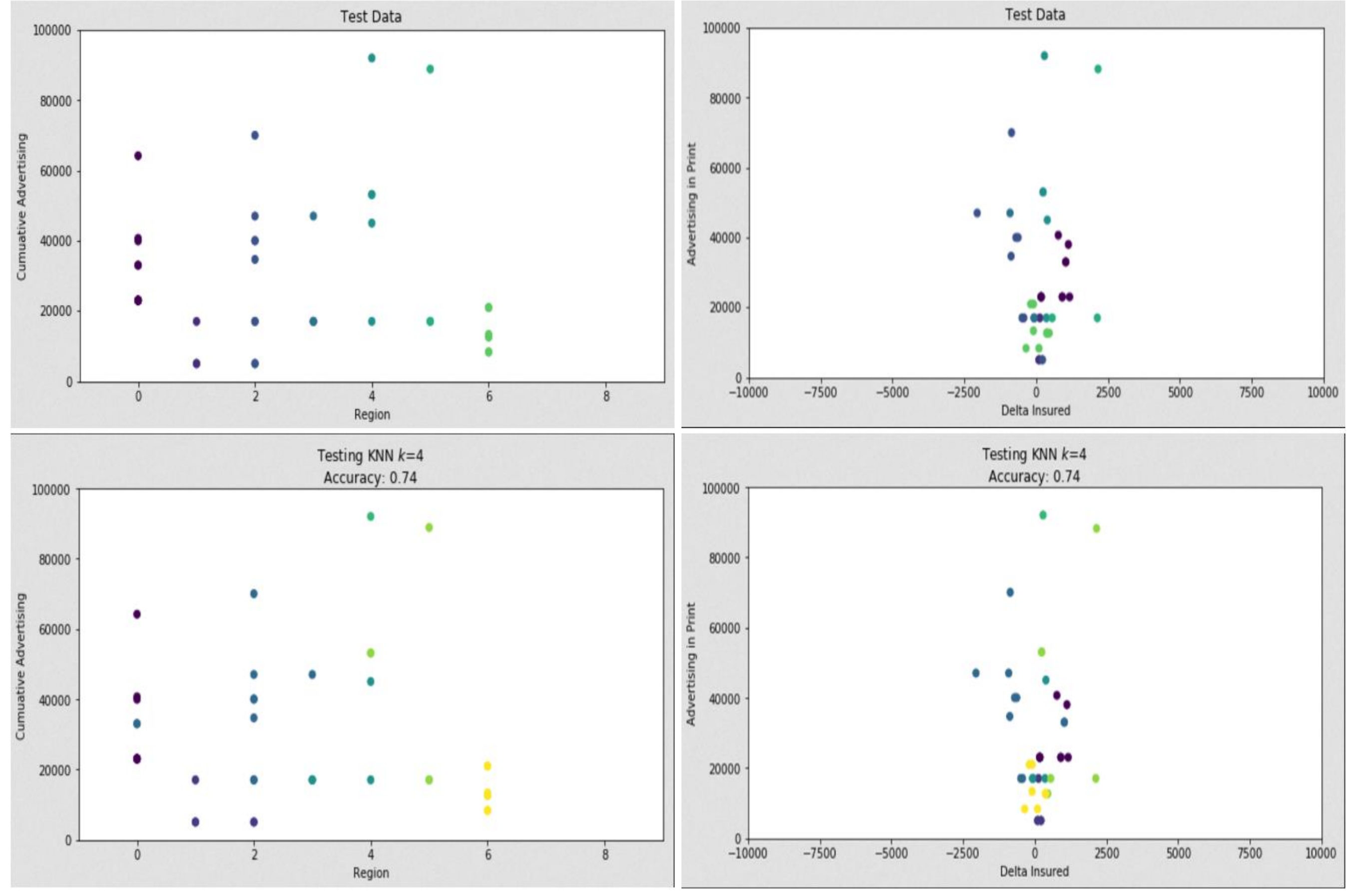
2. Another issue was the lack of coherence in the advertising data with the MHIS. The MHIS provided an extremely large amount of generic population data, but most of it was either too specific or too ambiguous for our feature space; the advertising data had few "links" to most of it. Specifically, for say something like broadcasting data, the "Impressions" data had no specific ratios or percentages that reflected a demographic, such as Latino, and thus something like a Latino population ratio from the MHIS could not be used due to statistical constraints.

Supervised Approach

In Data Science, a supervised learning algorithm is an algorithm that takes a set of data-points such as a large DataFrame and uses a specific set of "labels" or "IDs" that correspond to those data-points in order to make a predictive outcome on future data-points. Essentially, by using these "labels" like tags for a given row of data (any number of features may be present in this row of data, so long as it is coherent with all other rows of data), we can train an algorithm such that given a new row of data without a "label", we may be able to predict what that "label" should be.

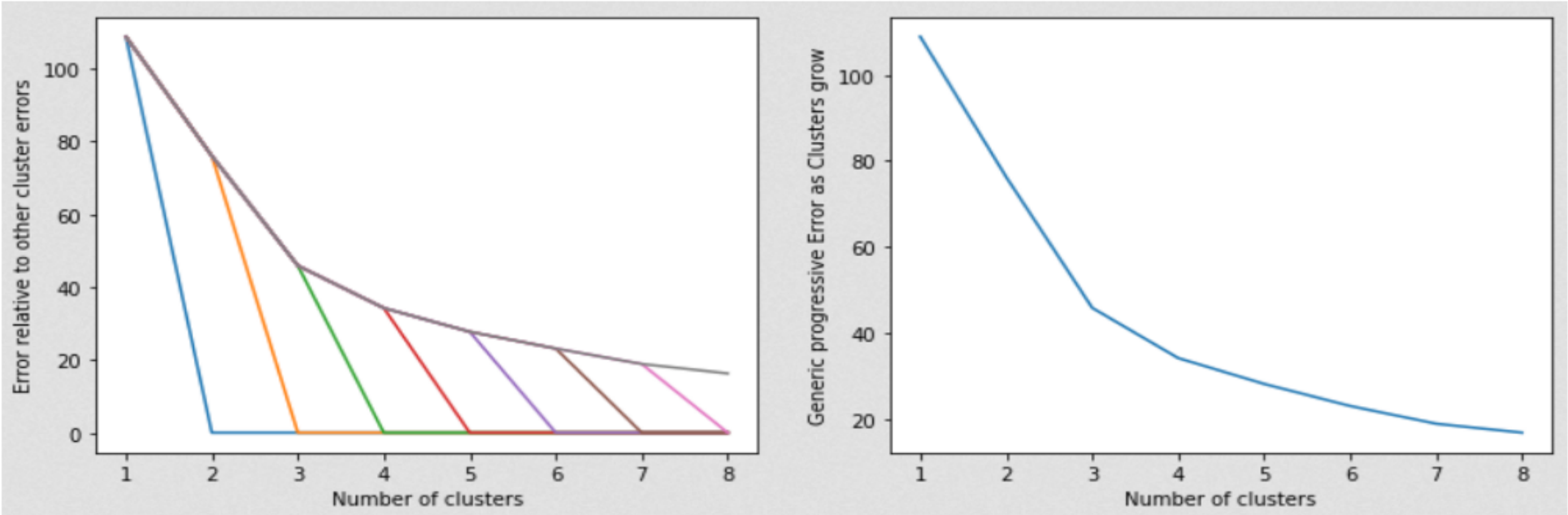
In the case of this project, we used a Regional feature to act as our labels in this approach. The MHIS provided us with at least 8 regions (Western, Metro Boston, Metro South, etc.) that could then be used with other cities that subset of these regions like Quincy or Mattapan. In conjunction with the advertising data, we checked how well the algorithms predicted advertising presence against regions.

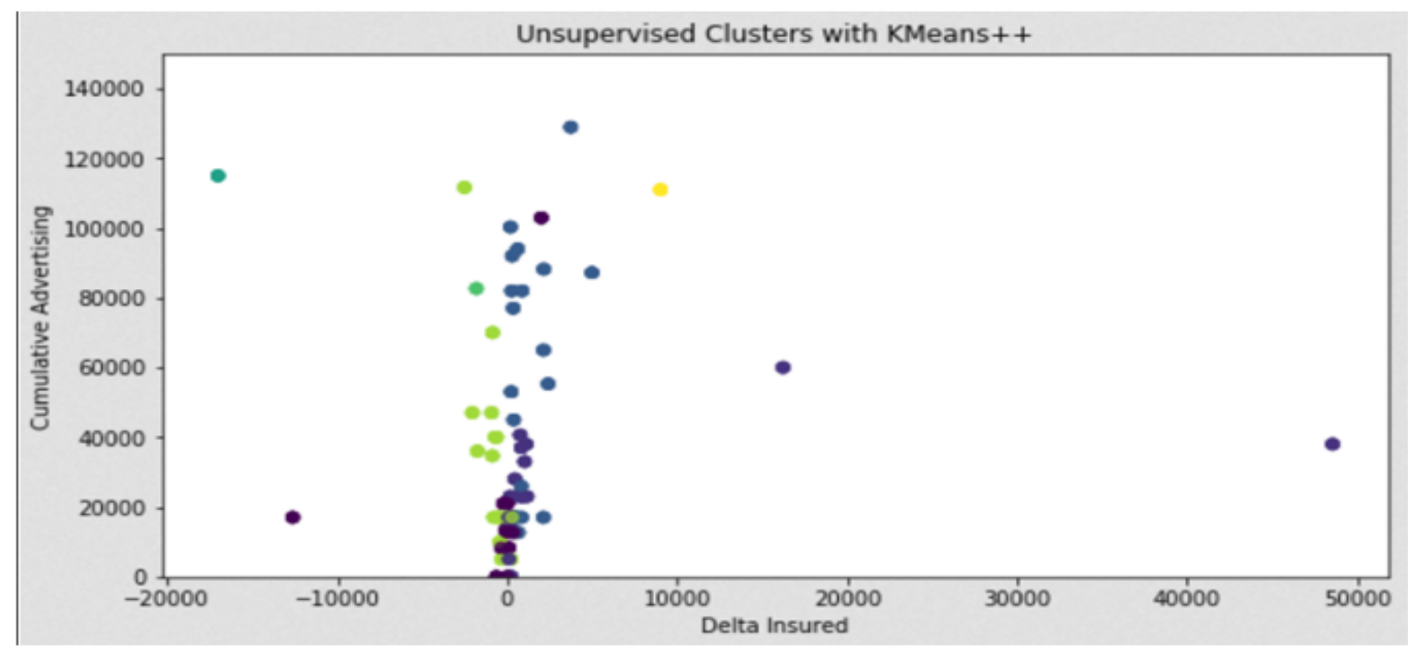
The most effective supervised algorithm used in this project was the KNearestNeighbors algorithm.



Unsupervised Approach

Another approach to analyzing data-points in Data Science is an unsupervised way, where all of the data-points in a large, DataFrame like structure are measured against one another for similarity (such as Euclidean distance) and then pooled into certain "clusters" based on their similarity. This is where the term "Clustering" comes from in this case. In our approach, we used a KMeans++ algorithm that measured the arithmetic mean of the data-points, among other things, using all features available. The corresponding graphs highlight the error in size of clusters used.





Conclusion

In conclusion, our current work on the data has exposed certain flaws. Firstly, there is a linearity in the data explicitly due to its geographic dependencies. This means our feature dimensions were limited by this and could not be "Regularized" against non-linearity. Additionally, the lack of coherence in data-points caused clustering and labeling to be, at times, sporadic. However, we did find that there was a relationship to certain regions and their change in insured individuals. It appeared these results were more regionally dependent and less advertising dependent.