MACHINE LEARNING AND VIRGINIA'S FOOD DESERTS

ML4VA CATEGORY: GOVERNMENT

Paige Hipes
aph3gx@virginia.edu

Mariam Guirguis
mg6qb@virginia.edu

Avni Garg
ag9wse@virginia.edu

December 9, 2021

1 Abstract

About 23.5 million people live in food deserts in the United States resulting in the USDA's efforts to further fund research to discover how and where food deserts come to be. The USDA defines food deserts as low-income areas that have limited access to nutritious and healthy foods. As a result, chronic illnesses have been positively correlated with food deserts, resulting in higher medical bills in areas classified as food deserts. Hence, a Virginia without food deserts is an improvement to public health– specifically working towards equity– as the majority of food deserts are home to minority populations. In order to research food deserts and aid in their elimination, different characteristics and features were classified, clustered, and regressed to allow for data discovery and a predictive tool to be created. The optimal classifier, clustering algorithm, and regression models chosen are polynomial SVM, Kmeans clustering with random initialization, and Random Forest Regression respectively. Clustering was used to visualize relationships between features and determine the geographic locations of food deserts in Virginia. Classification was used to understand the binary relationship between food deserts and different features. Lastly, a regression model was created and extensively hyper-tuned to aid in the creation of a prediction tool: a space for users to input demographic feature data into the model (either representing a hypothetical area or changes to an existing area) and receive a prediction of the share of of citizens living in a food desert. As a result, the Virginia government can proactively fight the issue of food deserts through machine learning.

2 Introduction

The United States Department of Agriculture (USDA) defines food deserts as low-income areas that have limited access to nutritious and healthy foods. Although the specific mileage varies definition by definition, food deserts are often quantified as low income urban areas further than 1 mile from a large grocery store or supermarket along with low income rural areas further than 10 miles away from a large grocery store or supermarket. According to a report from The Food Trust in 2015, over 1.7 million Virginians, including over 480,000 children, live in lower income communities considered food deserts, and this limited access poses several risks to public health. Those living in food deserts are more likely to develop diabetes, obesity, and cardiovascular disease. Resulting medical bills following disease may impact those living in food deserts as well. In order to address Virginia's food desert problem, it is important to understand the situation at hand; how different features affect a location's status as a food desert; and how to predict food desert risk in a certain location. A Virginia without food deserts is an improvement in public health and equity, as the majority of food deserts are home to minority populations.

Due to the seriousness surrounding the issue of Food Deserts, previous methods and studies were researched to determine previous work. The U.S Department of Agriculture conducted a one year study to assess the extent of areas with limited access to affordable and nutritious food and identify characteristics and causes of such areas. From the study, it was determined that existing data and research is insufficient to conclusively determine whether areas with limited access to grocery stores are classified as having inadequate access. However, it was determined that urban areas with limited food access are characterized by higher levels of racial segregation and greater income inequality. The research team used kernel density function to define low-income areas: the household income had to be below 200 percent of the Federal poverty threshold and for a given population, at least 40 percent of the total population had to meet the poverty criterion. Another study conducted in Montreal used hierarchical clustering to identify food

deserts in Montreal. After conducting their analysis, they came to the conclusion that proximity (distance to the nearest supermarket), diversity(number of supermarkets within a distance of less than 1000 meters), and variety (in terms of food and prices) are important in identifying food deserts. The research group performed computation of a hierarchical cluster analysis to classify and characterize census tracts (CT) in different groups of CTs with similar levels of the 3 measures aforementioned. The objective of the hierarchical cluster analysis is to locate food deserts and categorize all CTs in terms of deprivation and accessibility. Having similar data to the studies, we wanted to further their analysis by investigation of a new topic- whether demographic features of a land area can be used to predict the share of that population, classified as low income individuals, beyond 1 mile from the supermarket. In order to research food deserts and aid in their elimination, different characteristics and features were classified, clustered, and regressed to allow for a predictive tool to be created.

3 Method and Experiment:

The data set obtained consists of all US Census Bureau defined land "tracts" in the US with feature information related to demographics of the areas and levels of food insecurity. Wanting to localize our project to Virginia, we filtered the data to only include Virginian land tracts. The filtered subset includes 1,900 instances rather than the original 72,000. Preliminary analysis on the data set was performed to determine the nature of the pipelines needed. Since there are no missing values, no data cleaning is needed aside from feature engineering. Having over 100 features in the data set, we decided to use the high correlation filter to reduce the number of features. High correlation between two variables means they have similar trends and are likely to carry similar information. This can bring down the performance of some models, such as linear and logistic regression. Hence, using a threshold of a correlation of under 0.7, we reduced the number of features to 25.In addition, we removed 2 additional features that had the same value in every column. In addition, a low variance filter was created and run and the variance for the 23 remaining features was very low causing no additional features to be removed. Therefore, the next step is to split the data into the test and train sets and specific pipelines were created for each of the algorithms.

3.0.1 Clustering

The first learning approach used is clustering. We started off running k-means clustering on training data followed by Mini Batch K-means. In order to narrow down which k parameter (number of clusters) to specify, we created an elbow graph and found that the elbow occurs at k=2 (see Graph 1). For visualization, we plotted resulting clusters through each location's cluster label and longitude and latitude values (see Graph 2). For instance, when plotting the latitude and longitude of cluster label values (see Graph 2), it is interesting to see that one cluster (blue) exists in more dense and coastal areas, while the other (green) cluster exists in less dense and more western areas of the state of Virginia. Because we are dealing with geographical data, it is important to visualize such clustered relationships and understand them in order to move forward with deeper analysis. Concerning performance, we observed inertia values of resulting models: Mini Batch k-means (inertia of 212436.0027), k-means++ initialization (inertia of 210332.3469), and random initialization (inertia of 209854.3193).

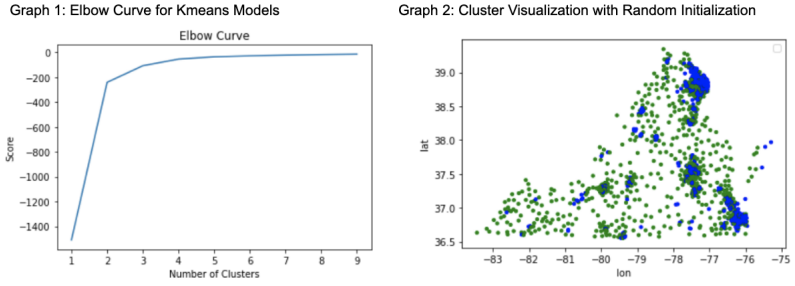

Figure 1: Graphs 1 and 2.

3.0.2 Classification

The next learning algorithm used is classification. We used 6 classification methods to try and classify areas as food deserts to try and get a better understanding of what type of model best fits the data. For instance, while the data is linear, it may be best fit by a Gaussian model or a combination of models. This is why it is important to test multiple classification strategies in order to find the best model that does not overfit and is generalizable to our dataset. We want

our classifier to be as generalizable as possible, as we are building a system that shows the prevalence of food deserts in Virginia which is a very geographically, income-based, and racially diverse state. We ran a number of classifiers on our training data to observe some results. We tried linear SVM (precision: 1, recall: 1), polynomial SVM (precision: 0.9623, recall: 0.9107), rbf SVM (precision: 0, recall: 0), K neighbors classifier (precision: 0.8929, recall: 1), random forest classifier (precision: 1, recall: 0.9643), and Adaboost classifier (precision: 1, recall: 1).

Based on the 6 methods tried on the train and test set, we find the polynomial SVM to be the best suited because of its robustness and the fact that it includes a quadratic term to better fit the shape of the data.

3.0.3 Regression

The last learning method used is regression. Four regression methods were used to try and predict the share (proportion) of an area classified as a food desert given certain demographic parameters. For instance, when a few features should be important, it is best to use elastic or lasso regression. Random forest regression was also used to cast a wider net through an element of randomness. To avoid overfitting because we want our model to be generalizable, cross validation with 5 folds was used. The following validation results from cross validation with 5 folds to avoid overfitting was observerd: linear regression (mean RMSE of 1.4043), random forest regressor (mean RMSE of 0.0471), elastic net (mean RMSE of 0.1501), and lasso (mean RMSE of 0.1501).

Based on these results, Random Forest Regressor was chosen as our classifier to move forward with. Next, we performed hyperparameter selection on Random Forest Regressor due to its low mean RMSE in cross validation. We found a GridSearch best estimator with n_estimators= 30 and max_features= 8 with a mean RMSE of 0.0622. We also found a RandomizedSearch best estimator with n_estimators= 190 and max_features= 7 with a mean RMSE of 0.0611.

Due to the low RMSE of the RandomizedSearchCV model, we moved forward with this best estimator and evaluated this final model on the test set, resulting in a final RMSE value of 0.0516.

4 Results:

Having chosen our classifier, regressor, and clustering to be Polynomial SVM, RandomForest, and Kmeans respectively.

For Polynomial SVM, we have precision as 0.9622 and recall as 0.9107. These metrics indicate accuracy and at the same time not overfitting the model to ensure generalizability.

For the RandomForest regressor through RandomizedSearchCV, the validation RMSE 0.08672100757801843.

For the Kmeans clustering algortihm, the inertia is 209854.3, which is the lowest of all the clustering algorithms we ran.

5 Conclusions:

After seeing the results of clustering, classification and regression, we decided to use regression to create a predictor that predicts with confidence how much of a population in Virginia is living in a food desert. We know that this is especially useful as living in a food desert is a public health hazard and should be rectified as soon as possible.

We can interpret our results to mean that we were able to successfully use machine learning techniques to gain insights about the predicament of food accessibility in the different regions of Virginia.

This predictor has the capabilities of enabling better city and rural planning in order to create a more equitable society, especially for minority populations as they are the most susceptible to be living in these areas. This way, Virginia can take the next step towards creating a more equitable society for its residents. There is potential for future work in this area by improving upon the regression and the predictor in order to plan communities better.

5.1 Limitations

One limitation we discovered is that we cannot plot SVM as the underlying grid of SVM is a 2 value cartesian co-ordinate system which cannot be predicted upon by our pre-fitted SVM which fits on a feature matrix of 171 columns. It also does not make sense to plot our dataset based on two aspects of our dataset as that would mean fitting on only two features which discards a plethoa of relevant information about what defines a food desert. Therefore we cannot plot SVM while having a representative model that classifies an area as a food desert.

6 Team Contributions:

All of us identified and agreed on a topic. We then divided the paragraphs and research. Paige did the motivation and dataset paragraph, Mariam did the Dataset paragraph. Avni did the intended experiments paragraph. We all checked each other's work and met again to go over it and make edits before submission.

We worked as a team to divide work for the checkpoint and also schedule regular meetings to check in, share progress, and review. Mariam completed data initial data investigation, visualization, and cleaning. Avni worked on classification models. Paige worked on clustering and regression models. Finally, we all worked together to synthesize and write up the results.

For the video, we all worked together to come up with the concept and execute it. Mariam and Avni wrote the script, while Paige filmed and edited the video.

For the final report, we all sat together and wrote the report while verifying its contents with each other.

Finally, we would like to acknowledge our mentor Alanna's contribution through her constructive feedback and encouraging comments. It motivated us to get through the project and deliver something we are proud of.

References

[1] Apparicio, P., Cloutier, M.-S., amp; Shearmur, R. (2007, February 12). The case of Montréal's missing food deserts: Evaluation of accessibility to Food Supermarkets. International Journal of Health Geographics. Retrieved October 7, 2021, from https://ij-healthgeographics.biomedcentral.com/articles/10.1186/1476-072X-6-4.

[2] ver Ploeg, M., Breneman, V., Farrigan, T., Hamrick, K., Hopkins, D., Kaufman, P., Lin, B.-H., Nord, M., Smith, T. A., Williams, R., Kinnison, K., Olander, C., Singh, A., amp; Tuckermanty, E. (1970, January 1). Access to affordable and nutritious food: Measuring and understanding food deserts and their consequences: Report to Congress. AgEcon Search. Retrieved October 7, 2021, from https://ageconsearch.umn.edu/record/292130/.

[3] Waldoks, R., Lang, B., amp; Treering, D. (2015). Food for Every Child: The Need for Healthy Food Financing in Virginia. The Food Trust. Retrieved October 6, 2021, from

`http://thefoodtrust.org/uploads/media_items/virginia-mappingfinal-lowres.original.pdf.`