Delving Into the Desert:

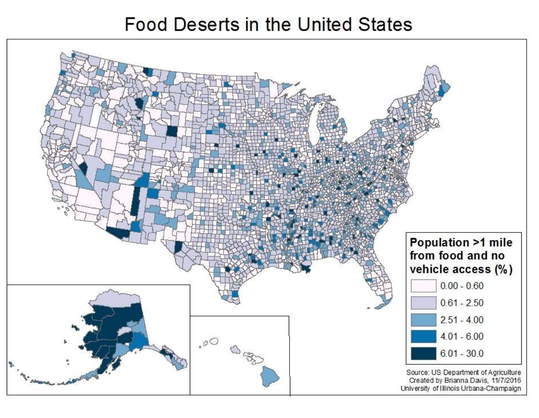
Using Deep Neural Networks to Predict Food Deserts in the United States

Maddy Mulder

Problem Statement:

11.1% of households (14.3 million) in the United States are currently food insecure, meaning they are without reliable access to a sufficient quantity of affordable, nutritious food. One feature of food insecurity is embodied by the phenomenon of food deserts where residents lack adequate access to affordable, healthy food. Currently, more than 23 million households fall within a food desert. Food deserts are a spatial phenomenon, driven by socio-economic factors that have resulted in a public health crisis. Food deserts are often found in low-income and minority neighborhoods, and these populations are left at a greater risk of developing health problems. The obesity rate is notably high in populations of color. This is often because they lack the financial means to afford healthy foods and the physical access transportation to stores with healthy food. Instead, they are forced to rely on fast food and convenience stores that have low-quality, unhealthy foods that compromise the health of those that rely on them for sustenance. The link between socio-economic and health status and food deserts is clear, but what if we could leverage what we know about this relationship to our advantage?

Using a machine learning algorithm, it would be possible to further explore and isolate the features of a community that are most likely to lead to the creation of a food desert. The algorithm could then use what has been discovered about the correlation between key features and food deserts to predict and highlight communities that have a high chance of developing into a food desert. Ultimately this information could be used by policymakers and community development professionals to identify key community characteristics that lead to the creation of food deserts and ameliorate the root causes of food deserts, thus moving from a reaction-focused response to food deserts to a more sustainable, and beneficial prevention-focused response to food deserts.



Data:

I primarily utilized data from the USDA Food Access Research Atlas (2015). This atlas provides food access indicators for 72,865 census tracts across the United States. There were a total of 148 indicators. For this project I chose to focus just on the data relating to the 1/2 mile for urban areas and 10 miles for rural areas demarcation zones. As such, I created my own Excel spreadsheet from the original USDA Food Access Research Atlas data where I narrowed down the original 148 indicator columns to just 25 for all states, including the District of Columbia.

Image Source: http://www.gpedia.com/en/gpedia/File:Food_Deserts_in_US_(2010).jpg

Models:

In the end, since this project was, at its heart, a binary classification problem I settled on creating two different models: a linear classifier and a boosted trees classifier.

Model One: My first model was a linear classifier (which is a logistic regression model). I used my feature columns as the argument for the classifier and trained it on the training input function I had created earlier. I ran this model three times, increasing the max_steps each time from 50 to 100 to 500.

```
linear_est = tf.estimator.LinearClassifier(feature_columns)
linear_est.train(train_input_fn, max_steps=50)
lin_result = linear_est.evaluate(eval_input_fn)
clear_output()
print(pd.Series(lin_result))
```

Model Two: My second model was boosted trees classifier as the goal of this project was to accurately predict whether or not a census tract was a food desert. I passed in my feature columns and a batch size of one for each layer as my arguments for the classifier and trained this model on on the training input function I had created earlier with max_steps set at 100.

```
n_batches = 1
est = tf.estimator.BoostedTreesClassifier(feature_columns,
                                          n_batches_per_layer=n_batches)
est.train(train_input_fn, max_steps=100)
result = est.evaluate(eval_input_fn)
clear_output()
print(pd.Series(result))
```

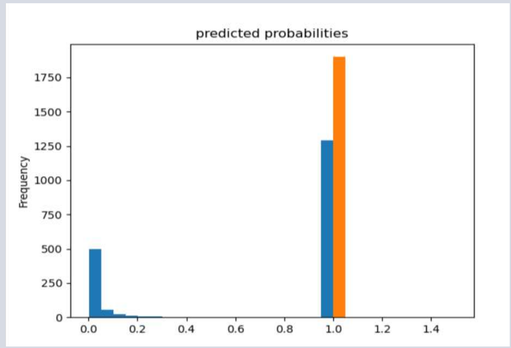
Next Steps:

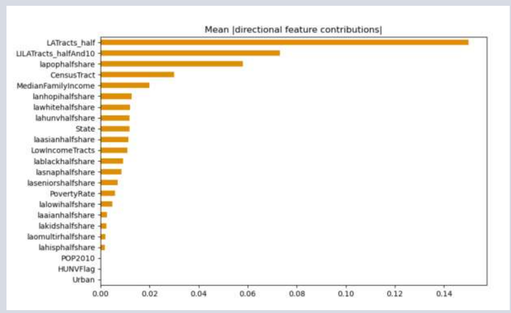
The project I had originally envisioned was much more expansive in terms of both data and model complexity. However, time constraints, the data I had access to, and my own newness to the world of machine learning forced me to scale my project down quite a bit. My smaller model thus left me with many potential ideas for next steps.

1. I focused only on data that related to low -food-access zones at 1/2 mile for urban areas and 10 miles for nonurban areas. The full USDA Food Access Research Atlas dataset also broke the data down into low-food-access zones of one mile for urban and 10 miles and 20 miles for nonurban areas and included data on how many households in each census tract have access to a vehicle. I think it would be a great next step to run my model on an expanded dataset that includes these data points.
2.Add the USDA collection of location data on farmers markets across the United States to my dataset.
3. Further expand my dataset to include health data. The Population Health Division of the Center for Disease Control and Prevention collects statistics regarding the overall health for census tracts across the United States. They have a project, called the 500 Cities Project, whose goal is to "provide city and census tract-level small area estimates for chronic disease risk factors, health outcomes, and clinical preventive service use for the largest 500 cities in the United States." Studies have shown that there is a strong link between the health of a population and the presence of food deserts I think it would be fascinating to use my model to examine this relationship in more depth.
4. Make this project a spatial inquiry. In the future this may be an interesting path to go down as food deserts are, at their core, a spatial phenomenon. For instance, including the spatial layout of transport networks within communities might be a key addition to this query as limited physical access to healthy foods is one cause of food deserts.

Results:

Model One (Linear Classifier):
On this model's very best run, which was with max_steps set at 50, I got an accuracy of about 69%, but the average loss was 1.732203e+08. Moreover, my the mean of the labels was .691 but the mean of my predictions was exactly one. Meaning this model always predicted that every tract was, in fact a food desert. As mentioned above I also ran this model with 100 and 500 max_steps, but let me tell you increasing the max_steps was the wrong thing to do!
Model Two (Boosted Trees Classifier):
This model faired much better, but I fear it is a bit overfit. It reached an accuracy of 98.8% and only had a loss of 0.035499. Moreover, its mean prediction value of 0.689561 was much closer to the actual mean label value of 0.690526.
Predictions
I trained both models to make predictions on the evaluation set (using the evaluation input function I had created earlier). The linear classifier always predicted a label of 1 for every census tract in Virgina, meaning it predicted that every tract was a food desert, when in reality only about 70% of the tracts were flagged as food deserts. The boosted trees model was much closer to accurately predicting the actual labels of each tract.





Concluding remarks:

Overall, the boosted trees classifier was much better, it had a much higher overall accuracy and a much lower loss. Additionally, its mean prediction value was much closer to the actual mean label value indicating that it was overall a more accurate model.

The LATract_half feature and LILATracts_halfAnd10 just indicated whether that tract was farther than 1/2 a mile or ten miles from a supermarket, so the feature feature that was in actuality most salient was the median family income. Various demographic features were also important. Most notably, the share of the population that was Native Hawaiian or Other Pacific Islander (lanhopihalfshare) or white (lawhitehalfshare) had considerable impact on whether or not the model designated a tract as a food desert.