Data Science #7
Los Angeles, CA

Project Overview

Farmers Market Analysis - Where can you sell your product?

This paper provides supporting information for a proof of concept (POC). The data product explained by this POC is a recommendation system focused on helping local farmers find the nearest farmers' market with the highest probability to sell their produce. The program could also help fans of farmers' markets to find the nearest market with product they are seeking.

1. Problem Formulation

If you were a local baker without a store front, how would you know where to sell your produce? What you if you had three farmers' markets within 50 miles of you that sold similar products but only one of them never has any stalls for baked goods? How do local farmers without a support system within the US agricultural infrastructure find their way to a farmers' market?

2. Create a Hypothesis

Create a recommendation system that allows the user of the program to determine what type of recommendation they are looking for. If they are buying, they are looking for the product sold in all of the "drivable" distances. If they are selling, they are avoiding saturated markets, long distances that may make the early morning drive costly, and successfully active farmers' markets.

Question related to this Data Science Project at time of Class Presentation: What County in the U.S. is mostly selling Wild Harvested produce or Cheese? And, vice versa?

3. Data Analysis & Testing (iterative)

User input:

Key parameters required by the programs' users would be location by means of Zip Code and County, desirable distance for travel and products of interest. Due to the population the program is intended to interact with, a web-front is needed.

Data Gathering:

The USDA maintains the National Farmers' Market Directory. This file is updated regularly because data input to the dataset is voluntary. Due to the frequency of data updates, it can be inferred that many farmers' market utilize this dataset for subsidies, marketing and/or advertising. (As such, it should be expected that this program already has a great deal of competition.)

Data Cleansing:

- Python allowed for recreating the 'updateTime'. Because the dataset is voluntarily updated, we can infer that the 'updateTime' is related to significant changes to the existence of the farmers' market. It can vary from initial establishment to creation of social media marketing to closing or relocation of the market. For the purpose of the POC, we assumed this date is mostly related to initial establishment so we can use Python and Pandas to visualize about how many farmers' markets are opening across the US by region, state or otherwise. Refer to https://github.com/sharooonie/farmersmarkets/blob/master/GA-Project-OverviewOfUSFarmersMarkets.ipynb for more information.
- Excel allowed for creation of a column called region. Python by way of MapReduce likely could have achieved the same results. Refer to https://github.com/sharooonie/farmersmarkets/blob/master/GA-Project-OverviewOfUSFarmersMarkets.ipynb for more information.
- The produce columns had inconsistent null values. Pandas facilitated the standardization of blank fields for modeling.
- See below for additional data cleansing activities that would further improve the quality and usability of this dataset.

Data Analysis & Test:

Model #1: Naive Bayes

Why use this? The original dataset was fairly clean. The amount of data cleansing required varied with the amount of analysis desired. As a result, the initial goal for modeling was to create clusters of information and utilize distance formulae to provide recommendations.

Successful or Not? This was not successful. While statistical metrics were obtained from running the models, the metrics were not the results intended. If a users executed this program, the program should response with an answer that resembles a value already existing in the dataset such as "Midnorth" or "Nashville". Statistical metrics are great for ensemble work but that was not the original intention either.

Model #2: Logistic Regression

Why use this? Logistic Regression is often used in epidemiology when testing pass/fail of viruses, vaccines, etc. Since the farmers' market dataset consisted mostly of binary values, this model was evaluated for suitability.

Successful or Not? This was not successful. Similar to Naive Bayes, the results were statistical metrics as opposed to predicted answers to so-called direct user questions.

Model #3: KNearestNeighbors

Why use this? The purpose of this model is intended to provide singular predictions. Such predictions are not statistical metrics, but are based on sound statistical distance calculations such as Euclidean distances.

Successful or Not? This was successful. When predicting the region most likely to sell Wild Harvested produce, the model output "Midnorth." From a business standpoint, the reasonable test checked. In further running this model for the region where the produce would not sell, where it could or could not sell by City, and then by County, all outputs passed the business reasonable test. Refer to https://github.com/sharooonie/farmersmarkets/blob/master/GA-Project-PredictWildHarvested.ipynb for more information.

Due to the success of Model #3, further tested with other produce. Refer to https://github.com/sharooonie/farmersmarkets/blob/master/GA-Project-PredictCheese.ipynb for more information. (Note that having prior business context for the data set evaluated greatly facilitated review of the success of models.)

Model #4: KMeans

Why use this? The purpose of this model is intended to provide singular prediction based on creation of clusters. Since the latitude and longitude for each farmers' market was provided in an interpreted fashion, the grid-like data can be used to facilitate location predictions.

Successful or Not? This was not successful (yet). Pandas successfully plotted all farmers' markets across the US in a visual shape similar to a map of the US. However, in validating the data values with Google Maps, confirmed the x and y provided are not literal latitude and longitude. Instead, they are pre-interpreted for a coordinate plot. As of the presentation for this project, the model has not successfully predicted the next x and y coordinates. However, I am optimistic this is the right model to complete the prediction. Refer to https://github.com/sharooonie/farmersmarkets/blob/master/GA-Project-ReviewLocations.ipynb for more information.

4. Create a Data Product

For the purpose of this Data Science project presentation, this paper contains the Python Notebooks and this README as a preview of the proof of concept. The data product itself will be built over time.

References Used:

1. National Farmers' Market Directory (USDA) :: http://search.ams.usda.gov/farmersmarkets/
2. Small Area Income and Poverty Estimates (US Census) :: http://www.census.gov/did/www/saipe/data/index.html

Additional Information:

The following are analysis conducted that were ultimately left off the final proof of concept.

(A) The US Census data for poverty and education were reviewed to help solve the original hypothesis. However, based on the unrealistic nature of a local farmer being able to sell produce across the country or oblivious to the economic conditions of surrounding livable areas, the hypothesis was edited to solve the problem. Additionally, the datasets appeared to already be results of previous analysis per review. The US Census datasets were therefore excluded from the remainder of the data science analysis.

LINK: http://www.census.gov/did/www/saipe/data/index.html</p>

(B) The social media data provided also allows for further prediction. For example, if a local farmer is looking to sell at a market where she can gain additional exposure for this farm, she may be looking for a market within traveling distance that does not already have the produce as a frequently sold item and that provides social media support on behalf of the farmers. As there are quite a few more inferences and analysis to confirm the inferences to complete, this analysis was started by not completed for this proof of concept.

LINK: https://github.com/sharooonie/farmersmarkets/blob/master/GA-Project-SocialMedia.ipynb

(C) The accepted forms of payment information in the dataset, like the social media information, could be very helpful for local farmers. Coupled with produce availability information, a local farmer could determine quickly if he needed to start a new farmers' market or join an existing one to further his business.

(D) The National Farmers' Market Directory records seasonal availability by market. However, the format of the data input is not easily interpretable. Python, Pandas, and their libraries could greatly improve the usability of the information in the dataset.