Predicting BMI from Food Acquisition and Availability

Alan Liu
CIS 3200-02 Data Processing and Analytics
Department of Information Systems
California State University Los Angeles

Abstract: The United States Department of Agriculture conducted[1] the National Household Food Acquisition and Purchase Survey (FoodAPS) in the period between April 2012 and January 2013 [1]. As part of the dataset, the various types of food sources within different radii were counted and recorded per participating household, as well as individual health data, including their BMI.[2] Using Microsoft Azure ML Studio, the dataset was cleaned and used to train Bayesian Regression models in various configurations. The best performing model yielded a mean absolute error of 2.039. While fairly accurate relative to the mean BMI of the dataset,[3] the low coefficient of determination[4] suggests the frequency of various food sources by itself is not the optimal predictor of individual BMI.

1. Introduction

The goal of this project was to accurately predict an individual's BMI given the frequency of food sources near their household. In the survey data collected by the United States Department of Agriculture, the participants self-reported their food acquisition events for the period between April 2012 and January 2013. In addition to this, socioeconomic and health data was collected from the participants. The participants' households were geographically referenced to each and every known food source, including grocers, supermarkets, superstores, convenience stores, fast-food and non-fast food restaurants. It is the count of these food sources that will primarily be used to predict BMI.

2. Related Works and Background

This project attempts to use the non-medical factors of an individual's food environment to predict their BMI. As such, it uses primarily uses factors relating to the accessibility of food.

There are other non-medical studies in finding contributing factors to BMI. Daniel Fuller et. al. published their study in The American Journal of Clinical Nutrition in January 2013 about the relationship between transportation mode, distance to food stores, fruit and vegetable consumption, and BMI in low-income neighborhoods [2].

The assumption was that using an automobile would convey easier access to food stores and would increase fruit and vegetable consumption, which would imply a healthier diet and lower BMI. Thus, those in low-income neighborhoods, being less likely to have access to an automobile, would have higher BMI because of presumably more difficult access to food stores and thus a healthy diet.

However, the study found no difference in fruit and vegetable consumption between different modes of transportation and found instead that those using public transportation had significantly lower BMI than compared to those who use an automobile [2]. This suggests that there are factors other than sole accessibility that impacts BMI, such as individual behavior and habits.

McFerran and Mukhopadhyay conducted a study on this behavioral aspect for predicting BMI. In their study, they studied people's personal beliefs on obesity and its cause. They assert that "such naive beliefs are important because they guide actual goal-directed behaviors" [3].

They found that people mainly believed that obesity is caused by either lack of exercise or by poor diet. Those holding the belief of the cause being lack of exercise were more likely to be overweight[5] than those who held the belief that the cause is poor diet [3]. They explained that those who held the low-exercise belief tend to consume more food than those who held the poor-diet belief.

From these two studies, we see that the factors contributing to BMI are complex and varied, leaning towards more behavioral factors than concrete factors. Nonetheless, the density of various food sources may still influence food-acquiring habits and affect BMI.

3. Platforms Used

The project was conducted using mainly two cloud-based platforms. The data cleaning, transformation, and the training/testing of the model were performed on the Microsoft Azure Machine Learning Studio, and some data were also exported into Elasticsearch's cloud service for visualization.

3.1 Microsoft Azure Machine Studio

Since Azure ML Studio is a managed cloud platform, the exact specifications of the underlying hardware are not made public. However, depending on the service level, the service provides different levels of service from its service. This project was done on the free tier.

[1] USDA's Economic Research Service and Nutrition Service co-sponsored the survey.
[2] Body Mass Index, calculated as weight (kg) / height (m)^2.
[3] The mean BMI in the dataset was approximately 25.
[4] The Coefficient of Determination was 0.596.
[5] BMI > 24.9 is categorized as Overweight; >30 is Obese.

In the free tier, the service is limited compared to the paid tiers. Storage space, used for saving the datasets used in the project, is limited to 10GB. Each experiment has a maximum of 100 models, experiments are executed 1 node at a time, and experiments can run for no more than 1 hour. This means for larger and more complex projects, the execution may not complete within the time limit.

3.2 Elasticsearch Service

Like Azure, the Elasticsearch service is also a managed cloud service. Unlike Azure, a platform may be provisioned based on the desired configuration. The billing is dependant on the configuration of the desired instance that is to be provisioned. The configuration used for this project is billed at the rate of approximately $0.02 cents an hour, with a monthly cost of around $16 dollars.

The configuration used is 'gcp.data.highio.1', whose underlying machine specification is 12 cores, 78GB memory, and 3000GB storage. The instance used for the project is configured to use 1GB of memory and 30GB of storage on the machine.

4. Data Preparation

The data used is a 181.3MB set of 'csv' files obtained from the official USDA Economic Research Service website [1]. There are several sets of data within but there are only a few that contain the data relevant to the project.

4.1 Food Access

For obtaining features relating to food access, 'faps_access_puf.csv' was used. The data in this file is keyed by household number, 'hhnum', and contains a number of columns describing the number of food sources within a certain distance of the household of a specific type.

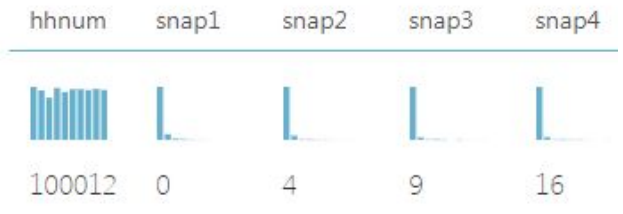

Figure 1. The first 5 columns of 'faps_access_puf.csv'

Figure 2. Breakdown of column names.

SNAP	SNAP-authorized retailers
SS	Superstores
SM	Supermarkets
CO	Combination Grocery
CS	Convenience Store
MLG	Medium or Large Grocery
FF	Fast Food Restaurants
NONFF	Non-fast Food Restaurants

Figure 3. Number component of column names.

1	within 0.25 miles
2	within 0.50 miles
3	within 1.00 miles
4	within 2.00 miles
5	within 5.00 miles
6	within 10.0 miles
7	within 15.0 miles
8	within 30.0 miles

For each column, the name denotes the type of food source, and the number denotes the corresponding distance away from the household within where the corresponding food sources are counted (e.g. 'FF5' meaning number of fast food restaurants within 5 miles). Each column name has 8 expanding radii, for a total of 64 features relating to food accessibility.

4.2 Individual Data

For obtaining the features and the future BMI label, the 'faps_individual_puf.csv' file was used. It is compositely keyed by household number ('hhnum') and the person number ('pnum') indicating the person within the household.

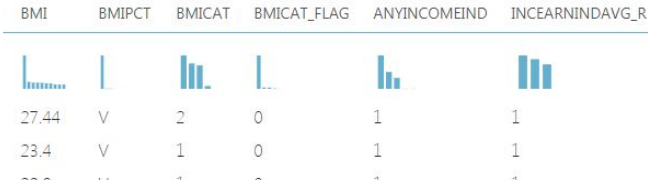

Figure 5. Example columns of 'faps_individual_puf.csv'

The columns contained describe various individual attributes of the associated survey participant. The column 'BMI' will primarily be used as the label for the model prediction.

In addition, there is a 'bmicat' column in the data with three values that denote the overweight status category: '1' denoting 'not overweight', '2' denoting 'overweight', and '3' denoting 'obese'. As a secondary objective, we will also try to classify this categorical column in a binary classification model — the column had been transformed with a Python script to change the 'bmicat' label to a binary class, '1' for not overweight, and '2' for overweight or obese.

4.3 Joining Food Access and Individual Data

Since both sets of data possess an 'hhnum' column, the two tables can be joined to produce a single table containing a row for every unique individual. Each row about the individual will possess the food accessibility features based on where they live — their household. The joining operation is accomplished using the Join Data module in Azure ML.

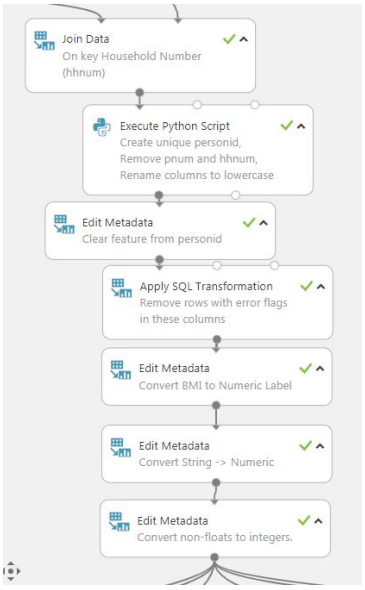

Figure 6. Series of data transformations.

After joining, a Python script was used to manipulate the tables to combine the 'hhnum' and 'pnum' fields into a single unique key by appending the 'pnum' to the 'hhnum' column[6]. The rest of the data cleaning was then performed, involving removing rows with errors and changing metadata, and the cleaned data was sent to the subsequent model training operations. Each model using this data splits the data into 66% training and 33% testing.

5. Results

In this project, several different configurations of prediction models were used. The various food accessibility columns will be the features used to predict the BMI label in a regression model. In the classification model, 'bmicat' will be used as the label, and the 'bmi' column will be removed to prevent it from influencing model performance since it should be reflective solely on the food accessibility factors[7].

5.1 Regression Models

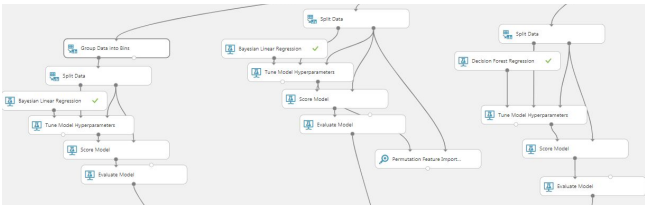

Figure 7. Binned and un-binned Bayesian Regression, and un-binned Decision Forest Regression.

[6] Scripts used available at the project Github repository: https://github.com/shozonu/usda-foodaps-bmi-model

[7] Using a Python script, also available in the repository.

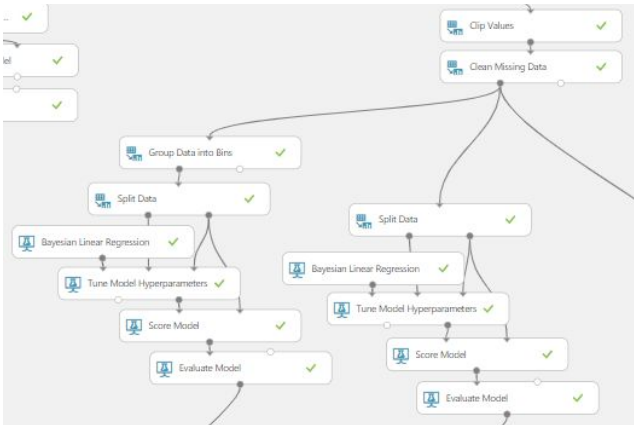

Figure 8. Binned and un-binned using values between 5th and 95th percentile.

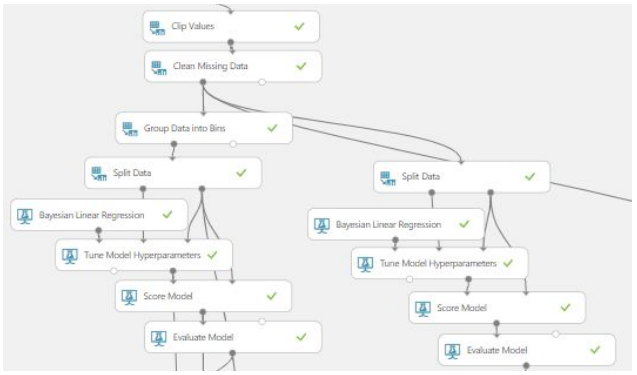

Figure 9. Binned and un-binned using values between 10th and 90th percentile.

Figure 10. Evaluation of regression models.

Type	Configuration	MAE	COD
Bayesian	Bin	3.3098	0.5605
Bayesian	No bin	3.3157	0.5607
Decision	No bin	3.2524	0.5734
Bayesian	Bin, 5-95%	2.3834	0.6061
Bayesian	No bin, 5-95%	2.3796	0.6061
Bayesian	Bin, 10-90%	2.0396	0.5959
Bayesian	No bin, 10-90%	2.0444	0.5951

The performance of the regression models is not very good as a whole. The low coefficient of determination across the board indicates that the label BMI is not very accurately predicted using the current labels. However, with average BMI being 25, a mean absolute error of approximately 2.0 is fairly accurate with the average error range being 7-10%.

5.2 Classification Models

For the classification models, the process is identical except that the model used is a Support Vector Machine model, and instead of the binning of features, cross-validation or lack thereof is used[8]. These models attempt to classify an

[8] The modules are set up the same basic way in Azure ML as the regression models, and are visually similar.

individual as overweight/obese or not overweight based on the food accessibility features used before, with the label being 'bmicat'.

Figure 11. Evaluation of Classification models.

Configuration	acc	prec	recall	fscore	auc
no clip, no X	0.622	0.633	0.875	0.734	0.622
no clip, X	0.615	0.626	0.871	0.729	0.627
5-95%, no X	0.620	0.632	0.909	0.746	0.585
5-95%, X	0.615	0.636	0.863	0.732	0.594
10-95%, no X	0.627	0.627	0.992	0.768	0.553
10-95%, X	0.615	0.622	0.952	0.752	0.577

Like the regression models, the classification models seem to have low performance. With an AUC hovering around 0.6, it is only slightly more effective than guessing. The abnormally high recall of the models compared to the lower precision and accuracy suggest that the model tends to over-classify rows as overweight/obese, due to the low number of false negatives (and consequently resulting in higher false positives), which leads to the low performance in the precision and accuracy metrics.

6.0 Visualizations

The dataset used in the models as well as other interesting peripheral data were imported in Elasticsearch and visualized.

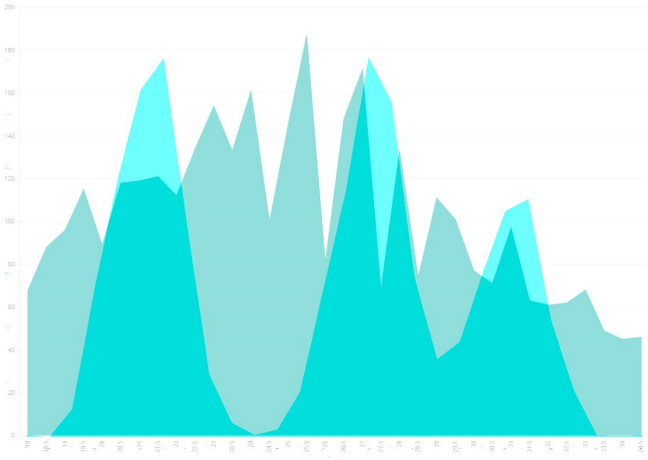

Figure 12. True BMI (darker) vs. Predicted BMI (lighter)

From the overlayed histogram in Figure 12, we can see that the distribution of BMI predictions from the model does not closely follow the distribution of actual BMI. The true BMI distribution seems to roughly be a standard distribution whereas the predictions seem to follow a polynomial pattern. We can see this evidenced in the large gap in the middle where the prediction fails to predict almost any BMI within the 23 to 25 range at all, despite the range being the most common in the dataset.

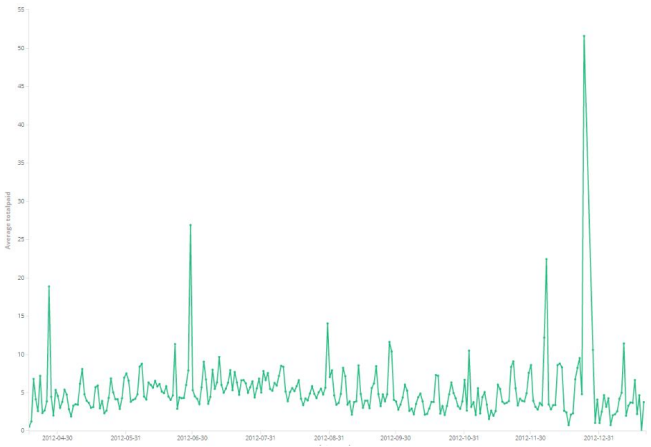

Figure 13. Time-series Graph of Average Money Spent

We can see the food spending behavior of the survey participants over a period of time. The regular spikes seem to indicate more spending on food during weekends, suggesting that either more people buy groceries on weekends, or more people eat out on weekends. The large spikes, especially at the end of the year, seem to coincide with holidays, such as Christmas.

Figure 14. Wordcloud: Eating Out

In Figure 14, we can see the various keywords relating to the food sources when eating out. Restaurants are common destinations, especially Texmex, Mexican, and American food.

The addition of frequent 'work', 'family' and 'store' keyword seem to also suggest that food-out events may coincide or may be in conjunction with other activities, such as store shopping with the family, or during work.

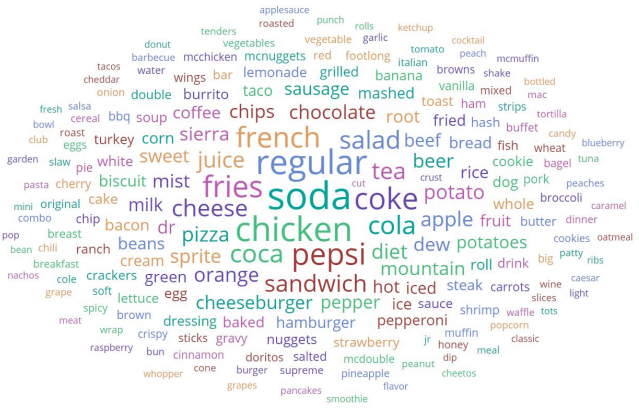

Figure 15. Wordcloud: Food Items

There are a wide variety of food items purchased by the survey participants. There are many terms that are frequented that are considered "junk food", such as soda and french fries. The frequency of the keyword 'chicken' may suggest, outside of restaurant contexts, that chicken is frequently purchased as a grocery item. This may be because of the relatively low cost of purchasing chicken compared to other meats.

7.0 Conclusion

Both the regression models tested and the classification models tested did not yield very high performance. The low coefficient of determination for the BMI regression model and the low accuracy yet high recall for the BMI category binary classification model suggests that there are additional factors relevant to the model are not present in the features.

Among the models tested, the Bayesian regression model with binned and 10-90% clipped values seems to be the most effective for predicting BMI, with a mean absolute error of approximately 2 — which, in this dataset where the average BMI is 25, is reasonably accurate for this purpose.

The classification models were a secondary objective, attempting to classify overweight and not overweight individuals. As such, they were not the main focus of this project, but the best performing classification model was the one configured for 5-95% clipped values and no cross-validation, with an accuracy of 0.627. However, the incredibly high recall (0.992) suggests this to be a byproduct of overly-aggressive classification of the 'overweight' class, meaning the model frequently classifies cases as 'overweight' when in reality they are not.

Because of this, its value seems limited, compared to the regression model's ability to produce BMI predictions within 7-10% margin of error.

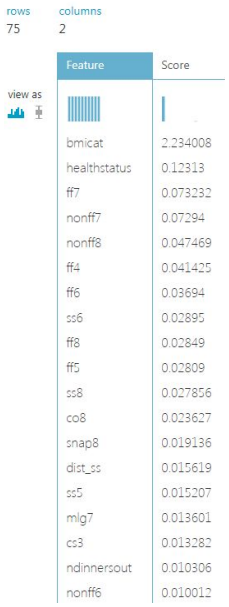

Figure 16. Feature selection for regression.

While there are multiple features positively influencing the performance, the current regression model seems to heavily rely on the 'bmicat' column, which presumably conveys a ball-park idea of the BMI. The 'healthstatus' feature, a self-reported rating of the individual's health, seems to also be an effective feature for predicting BMI.

Based on the feature importance, fast food frequency near a household seems to be among the next important features for predicting BMI. This suggests that food availability and access does indeed have some sort of influence on BMI. But considering the performance gap in the regression model, there is likely another factor that has not been taken into account.

Considering the results of previous works consulted, this seems to line up with the idea that there is some sort of behavioral factor and a medical-nutritional factor that is not being considered in this project's prediction of BMI.

References

[1] Economic Research Service, “FoodAPS Survey,” *United States Department of Agriculture*. 21 Feb 2017. https://www.ers.usda.gov/data-products/foodaps-national-household-food-acquisition-and-purchase-survey/

[2] D. Fuller, et. al., "Does transportation mode modify associations between distance to food store, fruit and vegetable consumption, and BMI in low-income neighborhoods?" *The American Journal of Clinical Nutrition*, Vol. 97, no. 1, pp.167-72. January 2013. https://doi.org/10.3945/ajcn.112.036392

[3] B. McFerran, A. Mukhopadhyay, "Lay theories of obesity predict actual body mass," *Psychological Science*, Vol. 28, no. 8, pp.1428-36, August 2013. https://www.ncbi.nlm.nih.gov/pubmed/23740551.

[4] Various Project Links:
https://github.com/shozonu/usda-foodaps-bmi-model
https://gallery.cortanaintelligence.com/Experiment/USDA-FoodAPS-BMI-Modeling