How Food Spenditure Affects Perceived Health: Stats 507 Project Report

Diana Liang Adam Stautberg

***Abstract*—As researchers continue to discover ways to improve the health of the general public, more emphasis is being put on prevention, rather than medication, of many conditions. In order to better understand how to prevent poor health, we must understand its causes. In this project we look at data from the NHANES related to money spent on food in order to determine the relationship between spending habits and health. Our model utilizes random forests to predict health, which is measured by the perceived health variable in the NHANES data, from multiple predictor variables relating to food spending habits. Our initial model had only 0.364 accuracy. We believed this to be due to the fact that 'good', 'very good', and 'excellent' health were very hard to distinguish between when survey respondents were answering their surveys. We hypothesized that if we separated the 5 categories into 'healthy' (good, very good, excellent) and 'not healthy' (fair, poor), our model would do much better. We saw this to be true, as our second model had a much better 0.785 accuracy. We concluded that these variables were good for a somewhat accurate assessment of overall health. However, they were far from perfect, and more specific predictor variables would go a long way towards forming a stronger model connecting money spent on food with perceived health.**

August 26, 2020

I. Introduction

A. Motivation: Food, Health and Money

Health is a great concern ranging from infectious diseases, which conclude with the eradication of the infectious agent, to chronic illnesses, which are ongoing often for an entire lifetime. While proper medication is crucial to health, prevention is becoming more critical. Chronic diseases in the United States have overtaken infectious diseases in the leading causes of death, and such staggering numbers have led experts to research these types of diseases in hopes of treating these conditions. So far research has shown that, while the exact causes may be more specific to the illness, an impressive number of leading chronic diseases have been linked to obesity, a condition that has affected more than 2/3 of the US population [1] and is proported to only increase. The American government has been trying to combat this rise by suggesting life-style changes dominated by physical exercise and diet changes to prevent these chronic illnesses, with an emphasis on the latter as more evidence has shown diet change to be more effective. And research shows that overall traditionally 'healthier' food options such as fruits and vegetables have linked to improving health [2]. The link between traditionally 'healthier' food options and improved health has become such common knowledge that suggestions of eating these foods seem more like repetitive nagging than actual advice. This ubiquity, though, introduces a slight contradiction. If it's common knowledge that eating healthy will lower risk of chronic diseases, why are so many people still dying from these chronic diseases?

The answer may lie in what people actually buy and what they actually consume. A transaction of money is almost necessary to obtain food, since almost every person obtains food from buying it. Individual food spending habits may not tell us much about the link between food spenditure and health, but greater trends and general transactions will give a less personal insight into the connection. The greatest transaction for food in the United States is from agricultural subsidies, most of which are corn subsidies [3]. Corn subsidies were connected with volume of growth in 1972 and, since then, the overproduction of corn has become a major issue [4]. This overproduction has inspired food companies to use cheap corn to produce food products. Easy examples include corn chip snacks, but more hidden products include foods with corn syrup. In addition, a majority of the corn produced goes to feeding livestock, which sickens the animals but lowers the cost of meat. While not every cheaper alternative may contain corn products, the trend of cheaper alternative foods containing corn products can still provide insight into the effects of cheap corn on eating habits and ultimately health.

If corn was nutritious, this seemingly omnipresent ingredient might not be a problem, but corn does not contain significant amounts and diversities of micronutrients compared to even the most common fruits and vegetables [5]. These subsidies artificially lower prices for these unhealthy choices to make them more appealing, which may be an applaudable marketing technique for food companies but have less than stellar implications for the health of their customers. Such foods also tend to have greater shelf-lives, allowing them to be transported to areas that do not have the infrastructure for constant delivery. This is why many lower income residences may be stuck in food deserts often still have access to non-whole food items. For those that do have access to traditionally 'healthier' foods, many choices that are better stocked may not be viable either. Some whole foods may seem cheaper than unhealthy foods, but their lacking in macronutrients and increased preparation time make them poor alternatives. Even with the USDA's tips for eating healthy on a budget [6], which focus on temporary price changes, the greater cost per calorie of healthy options makes it difficult for budget conscious consumers to make healthy choices.

Not having the funds to move outside of a food desert and to finance healthy choices impacts what foods people can eat. And since food is connected to health, having the monetary resource to have the option and then take the option of healthy foods is ultimately connected to the risk of many chronic illnesses. How people spend money on food should be linked to their eventual health, and we endeavor to explore that connection.

B. Analysis Set Up

Our main focus for this study is how money connects to health through food. If spending money on food affects the quality and quantity of food consumption, which eventually leads to overall health, how does this spending affect health? Specifically, which types of food spenditure affect health the most?

We've taken data from the National Health and Nutrition Exam Surveys (NHANES) on the CDC website, one of the most thorough ongoing national surveys of health available, to figure out what factors of spending are the most important, if at all important, in determining general health.

This is a difficult task with many hotly debated variables. One of the most accurate ways of measuring general health is actually perceived general health. While a doctor and sample tests do have applaudable accuracy in detecting and naming illnesses and health-worsening conditions, these resources only know about a fraction of a person's life and often need the person to notice symptoms before actually checking for a condition. While perceived health has a greater variance due to its subjective nature, it is comparable and sometimes more accurate in predicting general health as measured by lab instruments and statistics of mortality [7].

Our analysis will be based on creating a model using the Python scikit.learn package to accurately predict perceived general health based on factors connected with spending on food. The NHANES data includes the following variables that we will use as our predictor variables: number of meals not prepared at home, number of meals from fast food places, number of ready-to-eat foods, number of frozen meals, amount of dollars spent at supermarkets/grocery stores, amount of dollars spent on nonfood items, amount of dollars spent on food at other stores, amount of dollars spent on eating out, and amount of dollars spent on carry out/delivery foods. All of these factors have a maxed out quantity that does not accurately reflect the actual number or amount for these variables.

Since there was a high chance of missing values for every one of these variables, the response variable included, we combined data from the three most recent surveys with these variables of interest: the 2015-2015 survey, the 2013-2014 survey, and the 2011-2012 survey. And given the mix between continuous and categorical nature of the predictor variables, the best modelling technique for this dataset and our interest in the data is random forests.

C. Random Forests

Random forests are a powerful tool for prediction. They build on the already great qualities of classification trees to produce models that fit many types of data.

Classification trees create a chart of nodes and branches based on the most important predictor variables that split the response at the values that most cleanly separate the different classes of the response variable. Each node has a decision factor that classifies the data along different branches based on that decision factor. A classification tree can have many levels of nodes and branches that eventually lead to leaf/terminal nodes. These leaf/terminal nodes house the observations that fall into the classes of the response variable. The overall tree is then the model used to predict the response variable from the path the new data takes down the tree to the leaf/terminal nodes. What makes these trees useful is that the same predictor variable can be used at different levels of the tree and can be separated by different criteria along the tree as well. Based on these qualities, classification trees on their own can be powerful tools in creating prediction models [8].

Random forests use classification trees as a starting template and increase their accuracy by lowering variance [8]. A random forest is a collection of trees that are each created by selecting a fraction of the predictor variables as possible variables for each level. Each tree, then, selects only a fraction of the training data and provides an out-of-bag error estimate. This estimate is equivalent to the training accuracy for that model, providing insight into model performance without validation techniques. The combination of these individual trees provides a model that has even more advantageous properties. Some of these properties include: dealing with different types of data, inherently handling covariances between predictor variables, and providing a measure of importance among predictor variables. Unfortunately a disadvantage of random forests includes the inability to actually visualize the model due to the usually hundreds or thousands on trees.

We mainly chose random forests due to their variable importance measurement that would be a standard way to see how important each of the types of food spenditure is in classifying health, but other characterisitics make this an even more ideal fit for our specific questions. For our modelling requirements, a random forest model already has the advantage of being able to deal with likely heavily correlated and most likely not normal distributed predictor variables as well as our response that will hopefully not be equal in frequency between the different health statuses. On top of that this type of model will be useful for the variables that are usually continuous but may act categorical when the max value is reached. Since we are mainly interersted in the importance ranking rather than the exact measurements of how the variables affect health, the disadvantage of model visualization is not a concern.

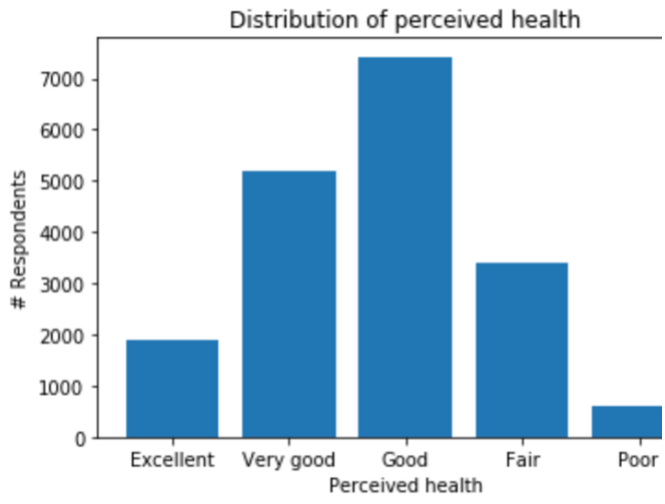

Fig. 1. Frequency of respondents for each category of health

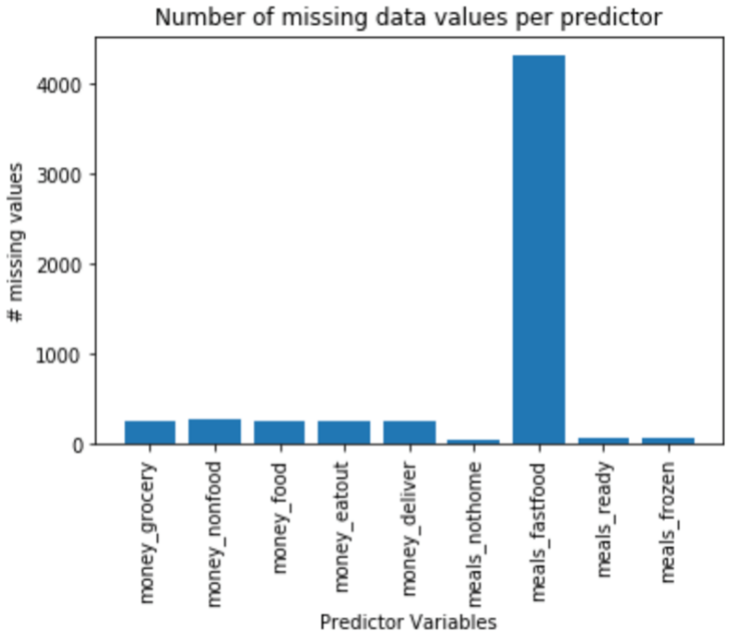

Fig. 2. Frequency of missing values for each predictor

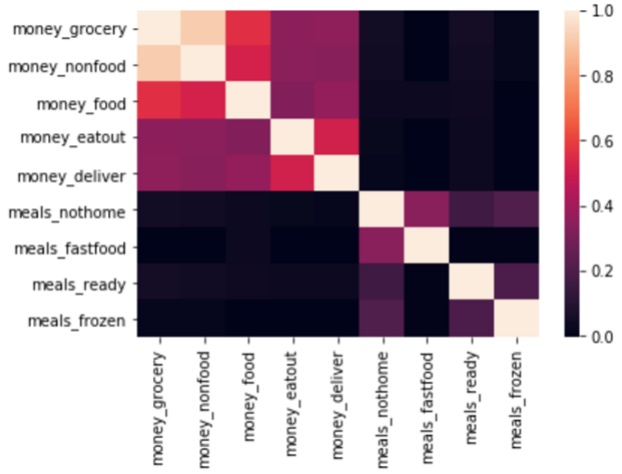

Fig. 3. Correlation heatmap between predictor variables

II. Predicting Health from Spending

A. Creating Data Set

Since the survey data was separated into different sections, but thankfully indexed by the same ID, we first needed to extract the columns that contained variables of interest (perceived general health as the response variable and the mentioned predictor variables) from the NHANES data sets for each year. And then each year's extracted dataset had to be combined together to create our data set. Luckily, Python's pandas package can store data sets as tables to be manipulated and merged, so we loaded each available table as a dataframe, extracted what we needed, and merged the pieces into a singular data set.

Next, we decided to also remove the observations where a health condition was not provided. These removed observations would be a nice illustration that our model can classify new data, but they could not help in judging model accuracy and would be extraneous to the purposes of our analysis. After this data preprocessing, the dataset was ready for data exploration and model fitting.

B. Data Exploration

In order to make sure that our analysis is meaningful we must first make sure our data set appears to be accurate. To do this, we focused on two main aspects of the data. First, what does the response variable look like? Figure 1 shows the distribution of responses to the question about perceived health. There are no outlying data points and the distribution looks to be about what you would expect, with most respondents saying their health is 'good'.

In addition to checking the response variable, we want to make sure all of our predictor variables relatively align to our expectations, as well as gain a sense of how many missing data points we have. There were no obvious outliers when looking at the spread of data for all predictor variables. Figure 2 shows the number of missing data points for each predictor variable. As you can see, most predictor variables, with the exception of 'meals fastfood', have relatively few missing points. However, 'meals fastfood' is missing 4314 data points, slightly less than a quarter of all respondents. It is unclear why so many respondents did not answer this question. Though we can still do our analysis, the fact that this variable has less observations is something to keep in mind when discussing our final results.

Due to our predictor variables being very similar, we figured that there would likely be a high correlation between some of the variables. Though our random forests should be able to handle this correlation, we still want to investigate to see how correlated some of our variables truly are. Figure 3 shows the correlation between each pair of variables. As

we can see, the variables are all somewhat correlated within two separate groups. The 'money' variables are all somewhat correlated with each other, having correlations ranging from .33 to .92. The 'meal' variables are less correlated with each other, with correlations ranging from almost 0 to .34. The highest correlation was .92 between the 'money grocery' and 'money nonfoods' variables.

C. Data Preprocessing and Model Tuning

The response and predictor columns were separated into Y for response and X for predictors and were again separated into train data and test data in a ratio of 8:2 respectively. This way we could measure how well our model would predict health status based on new data and overall give us a test error estimate to measure model accuracy.

The random forest functions in scikit-learn require a matrix without any missing values, which would be a challenge given the number of missing values found through our data exploration. As an expansive voluntary survey, NHANES datasets naturally contain a higher degree of missing data than preferred, so this is a common challenge. The best that we can do given the information from our dataset was to guess a value for those data based on an iterative method. While still experimental, this method makes predictions of these missing values based on values from similar observations. Most of our observations and predictor variables had a full set of values but we will have to consider how much the model relies on the number of fastfood meals as a predictor variable since most of the missing values came from that variable.

Then we trained a multitude of models with different numbers of trees and variables to the train dataset to figure out which parameters would provide the best fit. There was an array of number of trees (500, 1000, 1500, 2500, 5000, 7500, 10000) and number of variables (2, 3, 4, 5, 6, 7, 8) whose models we wanted to compared to find the best model based on highest training accuracy as measured by the out of bag accuracy. Since the computational time for all these models would have taken days, we decided to do a survey of each parameter on their own and then create a model based on a smaller array of the most promising parameters. The best candidates ended up being (1000, 5000, and 7500) for number of trees and (2, 3, 4, 5) for number of predictors. The out of bag accuracy for these different parameters are shown in Figure 4.

To achieve the optimal OOB accuracy, and by extension the expected model accuracy, the best parameters for the model were: 5000 trees and 3 predictor variables per node.

D. Model Fitting and Accuracy

The random forest was fit to the training data using the best parameters from model tuning: 5000 trees and 3 predictor variables, and this fitted model was used to predict health status from the test dataset and to calculate how accurate the predictions were compared to the test health status responses.

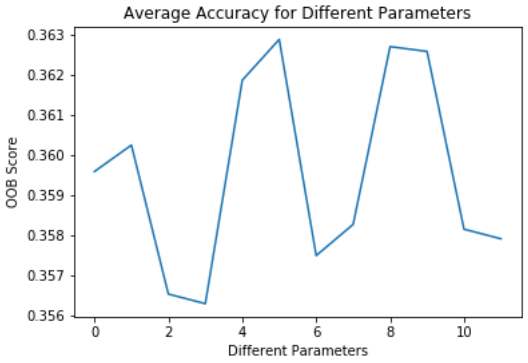

Fig. 4. Plot of OOB accuracy for train data

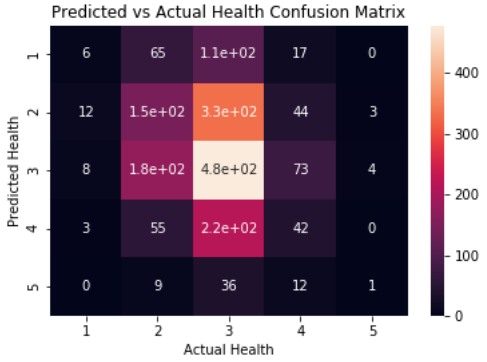

Fig. 5. Confusion matrix for general perceived health

With an accuracy around **0.364**, the model was fairly poor at predicting health based on how people spent money on food. This accuracy is better than random guessing, which would have been close to 0.20, but is not reliable in accurately predicting health. The main issue with prediction is the subjective nature of the classes. The classes for general perceived health are: (1) excellent, (2) very good, (3) good, (4) fair, and (5) poor. The distinction between excellent, very good, and good health is incredibly subjective, making it difficult for a human being to self-classify accurately nevermind a model. In the confusion matrix in Figure 5, a lighter color represents greater numbers of classifications. Many of the misclassification are from these middle health statuses such as (2) very good, (3) good, and (4) fair and the actual classification is usually adjacent to the misclassification. These most likely result from the vague distinctions what is considered generally good health since it would be incredibly hard to distinguish (2) from (3) compared to (2) from (5).

Even with these misclassifications, our model still faired better than random guessing. So there is likely some effect of how food spenditure on health that can be gleamed from this model such as looking at the variable importance plot to see which predictors were the most influential in determining health condition from how people spend their money on food, shown in Figure 6.

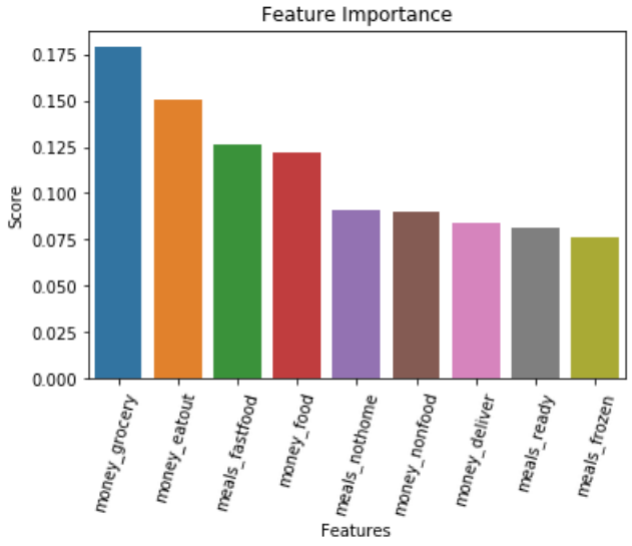

Fig. 6. Barplot of predictor variable importance

The most important seem to be the amount of money spent on groceries and the amount of money spent on eating out. The number of fast food meals is the next one after the two, but we will be cautious with interpreting whether this is meaningful due the number of missing values from this particular variable. In general, most of the variables about money seem to be more important than those about meals.

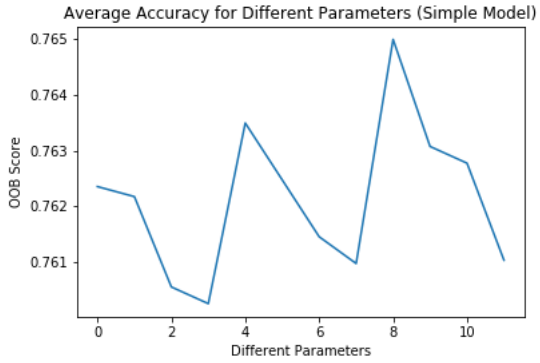

Fig. 7. Plot of OOB accuracy for simple model train data

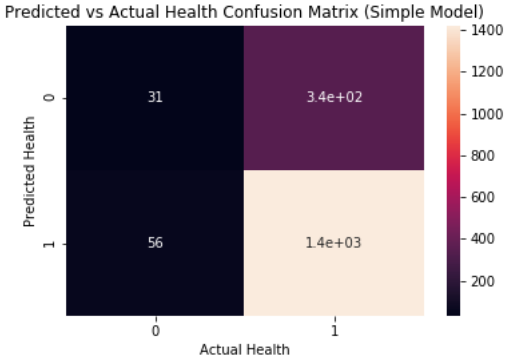

Fig. 8. Confusion matrix for simple model

E. Simple Model

Since we suspected that many of the misclassifications from the previous random forest model were due to the vague nature of the classes, we created a simpler model by grouping the 5 classes for health condition into 2 classes: healthy (1-3) and not healthy (4-5). Fewer classes would increase model accuracy inherently because the chances of randomly guessing correct have increased from 0.20 to 0.50, but exactly how much the accuracy of this simple model improved over 0.50 would provide a better understanding of whether food spenditure measured by these variables even had an effect on health.

We created a new binary variable that described our new classes and redid the data preprocessing for model tuning. Again we did a survey of different number of trees (500, 1000, 2500, 5000, 7500, 10000) and variables (2, 3, 4, 5, 6, 7, 8) as the tuning parameters and chose a smaller array of trees (1000, 5000) and variables (2, 3, 4, 5) to find the model with the greatest OOB accuracy. The plot for OOB accuracy in Figure 7 indicates that the parameters for the best fitting model were with 5000 trees and 2 predictor variables. The parameters for the model with the greatest OOB accuracy were chosen for the simple model, and the accuracy of that model was calculated from difference between the predicted and actual health conditions of the test data.

The accuracy of **0.785** dramatically improved from that of the original model as well as from the random guessing accuracy of 0.50. This much greater gap in accuracy between random guessing and the simple model is a better indicator that these predictor variables describing food spenditure do have an impact on the health compared to the gap between random guessing and the original model. Since this model can only account for 0.785 of perceived health status, though, this would probably not be the best model for predicting health.

Now that this simpler model is not as affected by the subjective nature of self-diagnosing health status, the resulting confusion matrix and variable importance rankings should be more accurate. The confusion matrix in Figure 8 shows us that, while the percentage of accurate predictions did increase, there are still a significant number of test cases that were misclassified. On the other hand, the comparison of predicator variables shown in Figure 9 is still very similar to that of the original model, hinting that the rankings were accurate despite the lower model performance. Again the most important variables were the amount of money spent on groceries and the amount of money spent eating out, followed by the number of fastfood meals. And in general the variables dealing with money were more important than the ones dealing with meals.

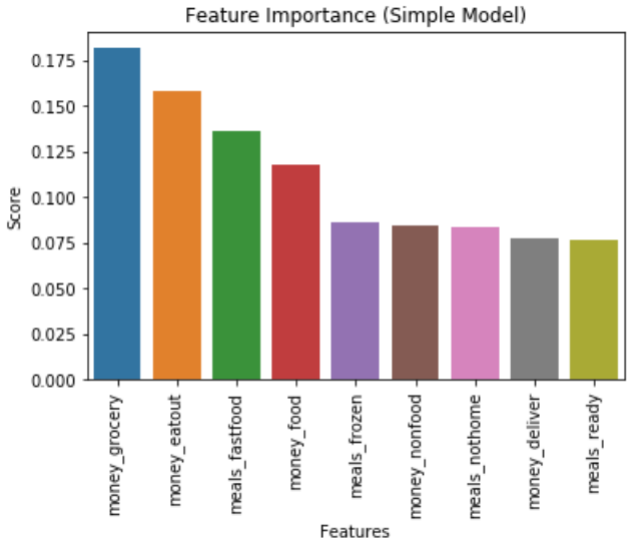

Fig. 9. Barplot of predictor variable importance for simple model

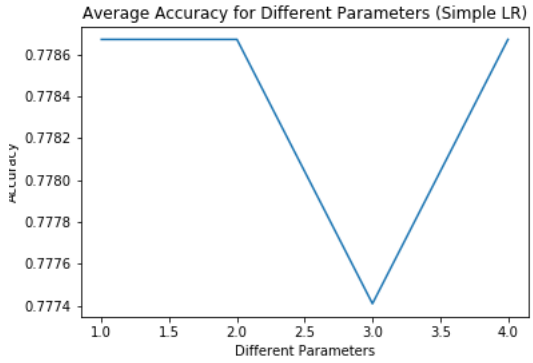

Fig. 10. Plot of CV training accuracy for simple logistic regression model

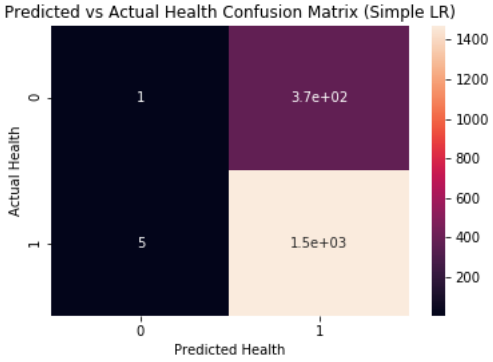

Fig. 11. Confusion matrix for simple logistic regression model

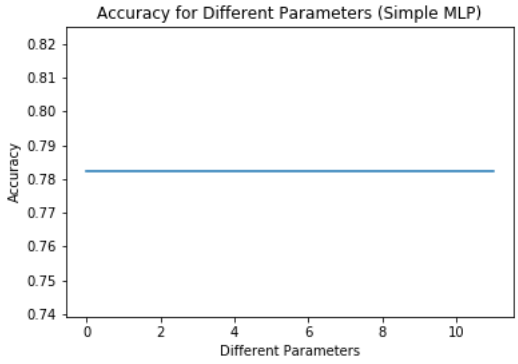

Fig. 12. Plot of training accuracy for simple MLP model

F. Logistic Regression and MLP

Since our simple model performed much better than the original model, we tried two other types of models to see if our model can be improved before analyzing our results. The other models should follow similar traits to random forests so that they can be more easily compared to the random forest simple model. Thus, we chose to do logistic regression and multi-layer perceptron to run on our simplified data.

Binary logistic regression is a common technique used for mixtures of categorical and numerical data that makes no assumption of normality on the variables, although it does assume that the relationship between the predictor variables and the log-odds of each class are linear. Despite that assumption, logistic regression is still a common and powerful tool for classification. For this model, the same techniques for transforming the data set from 5 classes into 2 classes was done. A binary logistic regression model from scikit-learn was tuned on the inverse of the regularization strength C for the training data. Using a CV of 5 folds to produce an average training accuracy, the graph of which is shown in Figure 10, we chose the model with the highest accuracy and fit that best model to the entire training data. The resulting test accuracy was **0.796** and provided the confusion matrix shown in Figure 11. Given the chosen parameters, interpretation of the relationship between health and the predictors does not make intuitive sense, so we focused on the model performance rather than interpretation of relationships.

In recent years, multi-layer perceptron models, or MLPs, have become more popular due to its high performance for complicated data sets and analyses. They classify using a multitude of hidden units and layers of unknown features without making any assumptions about the original data. While this simplified dataset only has 9 predictor variables, this method should still provide a model comparable if not better than the one for random forest. So the data set was simplified using the same techniques as for the simple random forest model and fit to by a multitude of MLP models, the training accuracy of which is shown in Figure 12. It was strange how fast the models converge and how the accuracy was the same for each, so the model was chosen at random to better understand the results. As can be seen in the confusion matrix of Figure 13, the MLP model predicted that every test observation was (1) healthy and none were (0) unhealthy and still managed to get a test accuracy of **0.798**.

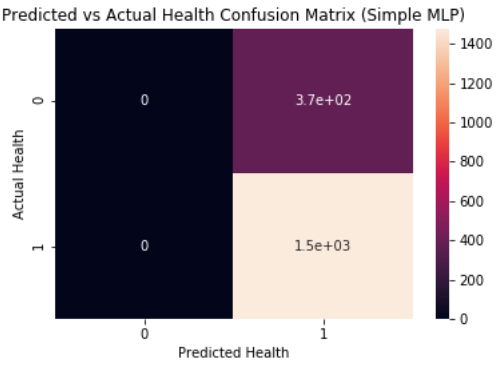

Fig. 13. Confusion matrix for simple MLP model

While the similar test accuracies between all three models suggest that the simple random forest model was performing to an appropriate degree given the data, it was shocking to see that simply guessing that every observation was (1) healthy would be the most accurate.

III. Analysis and Conclusion

A. Model Accuracy

Greater model accuracy is coveted because it can better describe the true relationship between variables and, thus, allow for decisions based on those relationships to have a more precise impact. Lowering the number of classes for health condition from 5 to 2 already dramatically increased model accuracy, but a model with an accuracy just shy of 0.8 is definitely not the best for describing how food spending affects health. In addition, the fact that the best performing model of the three simple models predicted the same class for all the observations brings the validity of all the models into question.

Since most of the predicted health statuses were (1) excellent, (2) very good, and (3) good; the distribution of (1) healthy and (0) unhealthy observations was drastically unbalanced and shifted towards (1) healthy. Thus, it's easy to guess that every observation is healthy and be mainly correct. That the MLP model could not do better than indiscriminate prediction suggests that these predictor variables are not correlated enough with health status to properly cluster by class. Yet both the logistic regression and simpler random forest models did attempt to classify some of the test observations as (0) unhealthy and still managed to achieve similar test accuracies, so there is a chance that food spenditure does have an effect on perceived health status, but the current data does not properly describe that relationship.

Overall these models tell us that there are major issues with the data that need to be addressed before making any definitive judgements on how food spenditure affects perceived health status. All the models achieved similar test accuracies but suggest diffferent conclusions about the relationship itself. For now, it can be assumed that there probably is some connection between the response and the predictors but that other predictor variables are necessary to fully understand what that connection is. Therefore it is still value in interpreting the variable importance rankings.

B. Which Spending

Since the logistic regression and MLP models do not have easily interpretable information about variable importance, we will focus on the information available from the random forest models. Both the original and simpler models had very similar measures and exactly the same rankings for predictor variable importance, shown in Figure 6 and Figure 9. While the model can definitely be improved with additional variables and clearer definitions of health condition, this pattern of predictor variable importance seems to be true for these specific predictors. This means that the predictor the greatest influences on health are the amount of money spent at grocery stores, the amount of money spent on eating out, the number of fast food meals, and the amount of money spent on food.

Interestingly, the distribution of variables related to the type of meal suggests that the type doesn't matter as much, so long as that meal is not a fast food meal. Even more interesting that the number of fast food meals is a more important variable than many despite its high frequency of missing values as displayed in Figure 2.

It's possible that the model did not perform better due to its reliance on fast food meals. Since almost 1/4 of respondents did not provide a value for fast food meals, the values were generated by an iterative prediction method that may not have accurately represented how many fast food meals these respondents actually bought. Either way a survey with more positive responses for fast food meals could change the overall structure of the model and its ranking in predictor variable importance. Some of the variables may also change ranking as a result, but the general distribution shouldn't change too dramatically. For now any comment on the importance of fast food meals on perceived general health based on these models is likely inaccurate.

As for the variables that aren't numerated by meals, the amount of money spent at grocery stores and the amount of money spent eating out are more important than the general amount of money spent on food, suggesting that where the food money is spent has an effect on perceived general health. The variables dealing with the amount of money spent were more important in general compared to the meal variables, suggesting that the effect of food spenditure on health has more to do with being able to spend money on food rather than the types of meals that were eaten. This could lead to connections between money and health that are not as directly linked to food. Factors such as socio-economic status and access to health care may be more influential in accurately predicting health status than these predictor variables.

To summarize, the amount of money spent in different establishments tended to be more important than the number of different type of meals. It has to be noted as well that many of these predictor variables are going to be connected.

For example, if someone is eating many ready-to-eat and frozen meals, they will likely be spending that money at a grocery store. It's not surprising that the variables amongst themselves have similar variable importance scores.

C. Future Improvements

There are still ways that we can improve future models looking into the relationship between food spenditure and health. The main improvement is to improve the data. Since the labels of perceived health status were difficult to differentiate, responders were most likely unsure which category was most accurate for their experiences. Having more discrete perceived health statuses would improve the ability for responders and the model to distinguish whether what relationship food spenditure has on health. In addition, the model could be improved with more observations for bad health. The lack in representation of bad health increased the difficulty of find features amongst those with bad health that would be useful in classification. While going further into the past surveys may seem like an easy fix, these other data may provide an inaccurate representation of the current relations not to mention the previous problem with vague health statuses.

More and different predictor variables will be necessary to both describe the food spenditure habits of Americans and more fully represent the how those habits affect health. Many of the predictor variables were correlated with each other, decreasing the necessity of all these variables, and together were not enough to provide a solid understanding of if and what relationship they had with health. As suggested by the variable importance rankings above, more variables should be added that are related to how much Americans are spending on different types of foods and in different locations such as how much is spent on whole foods at a grocery store versus how much is spent on processed foods at grocery store versus how much is spent on whole foods at a farmer's market. These more detailed accountings would be a more accurate depiction of food spenditure and hopefully provide a more understandable model.

While there's little that can be confirmed from the models in this study, there are still some worthwhile insights. There are a myriad of ways to improve the model to try to find some connection between food spenditure and health. Yet there could be little connection to begin with. The most important variables tended to be ones dealing with the amount of money spent. This could easily be a result of income and socio-economic status being highly influential in health, which implies that being healthy is much more difficult than spending the time and money to make healtheir food choices. As much as governmentally sanctioned food pyramids and health based blog sites want to believe that simply investing in healthy food choices will lead to health, it may be much more complicated.

IV. Contributions

Diana Liang wrote the introduction and conclusion as well as both the code and paragraphs for the data set creation and the 4 models analysis. Adam Stautberg wrote the code and paragraphs for data exploration as well as the abstract.

REFERENCES

[1] CDC, "Faststats - leading causes of death," March 2017.
[2] HHS, "Importance of good nutrition," January 2017.
[3] EWG, "Ewg's farm subsidy database."
[4] C. Frump, "Up to our ears: Corn overproduction, its environmental toll, and using the 2012 u.s. farm bill to limit corn subsidies, increase environmental protection incentives, and place accountability on crop operations," 2013.
[5] G. Goetz, "Ag subsidies fund junk food, report says," July 2018.
[6] USDA, "10 tips: Eating better on a budget."
[7] R. Heuberger, "Perceived versus actual health and nutritional status: Results from a cross sectional survey of rural older adults," *Journal of Gerontology and Geriatric Research*, vol. 3, no. 1, 2013.
[8] L. Breiman and A. Cutler, "Random forests."