Modeling Food Serving Sizes Through Nutrition Profiles

Jeremy Ondov

August 27, 2019

Project Introduction

Problem Statement

- Creating a new food product is a long and complex process
- Serving sizes incorporate many aspects of the product
  - RACC guidelines
  - Nutrition, cost, demographic
- Predictive model can give early estimate for servings

Data Sourcing

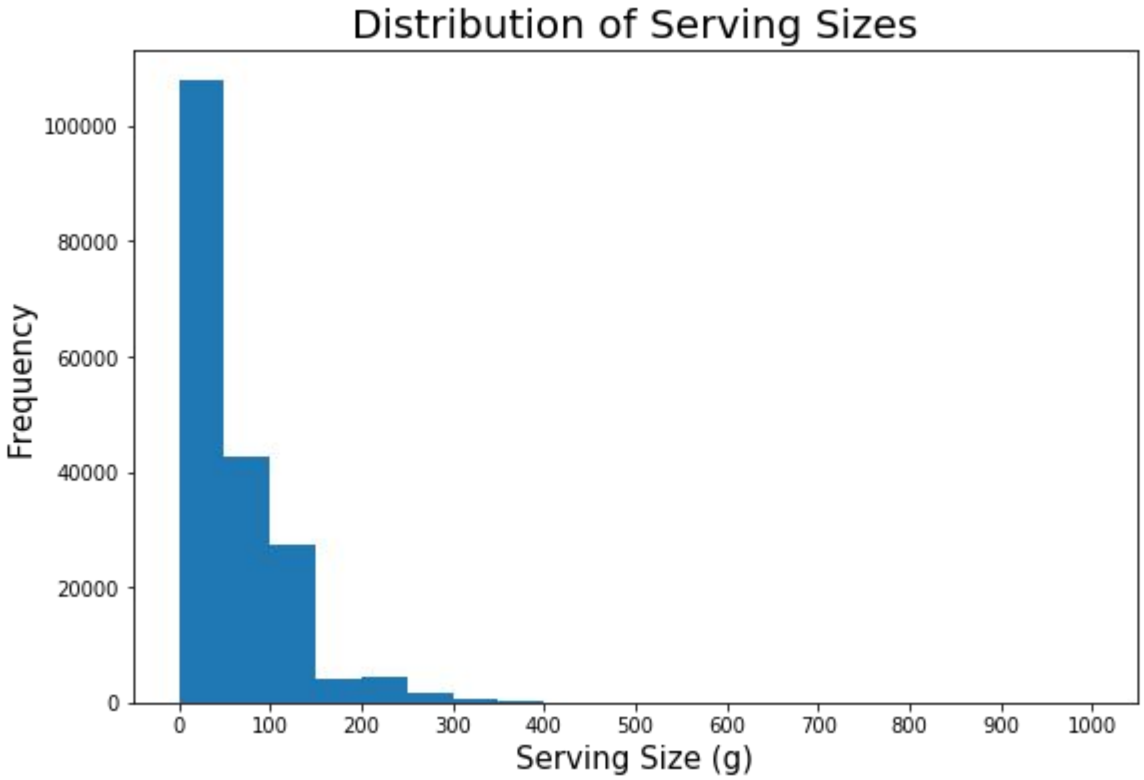
- USDA hosts a database of branded food products
  - Contains over 260,000 products
- Incorporates standard information
  - Ingredients, nutrition facts, market category

Data Analysis

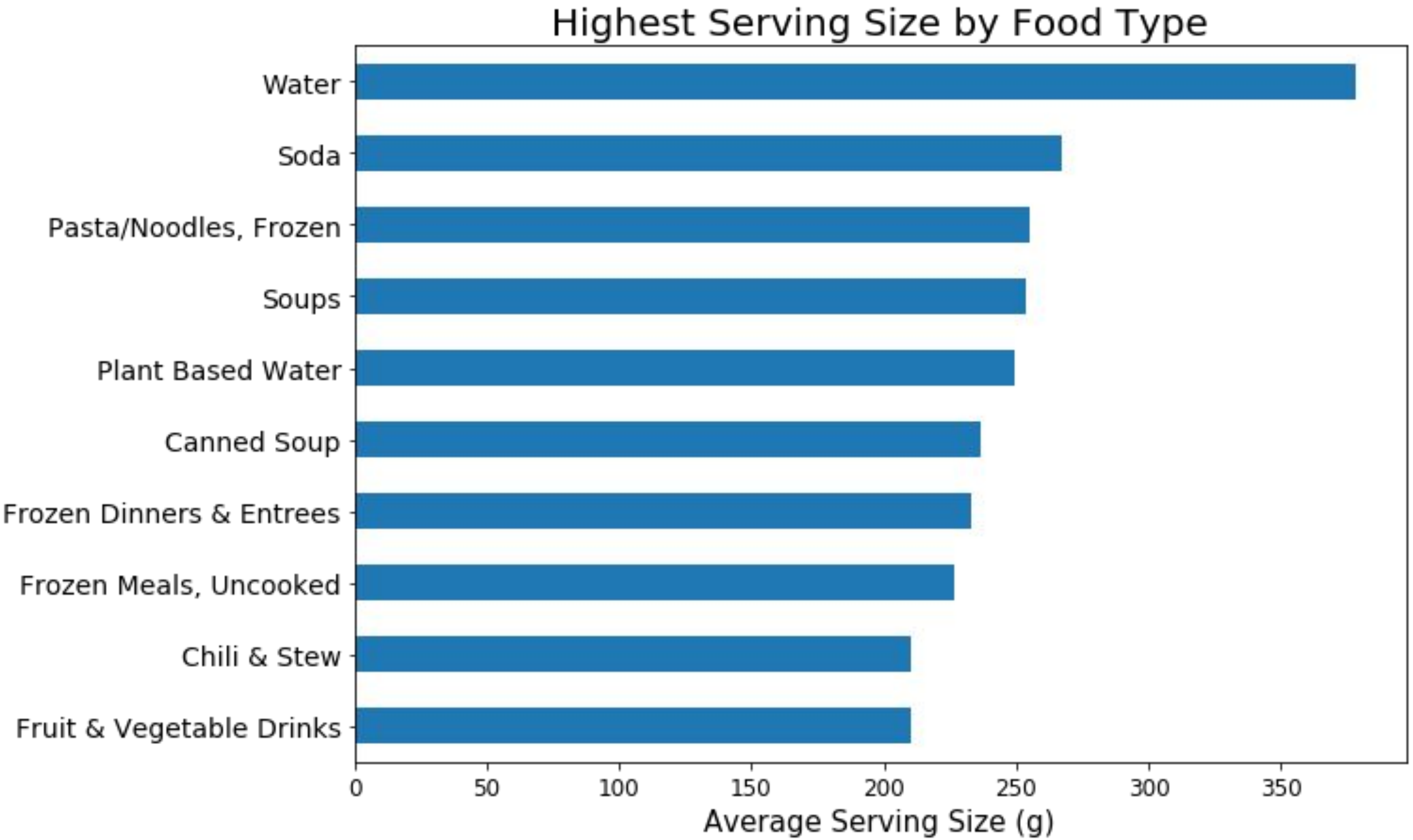
Cleaning Process

- Labels are not all the same
  - Standardized, with wide variations
- Many nutrient entries were missing
  - Macronutrients were used to estimate calories
  - FDA guidelines helped with imputation
  - Top 10 nutrients were kept

Serving Size is Right-Tailed



The largest servings are liquids and frozen meals



Dealing With Data

- Models trained on log-serving size
- Nutrition columns used as-is
- Product categories were dummied
- Ingredients were analyzed with NLP
  - Not included in final modeling at this time

Modeling

Model Types

- 3 Linear models were used
- Straightforward OLS fit on unscaled data
- Ridge and LASSO fit on scaled data
  - Utilized gridsearch and CV for best alpha

Feed Forward Neural Network

- Custom grid searching function was created
  - Altered parameters for given layers
- Final model had 3 hidden layers
  - 256, 128, 64 nodes
  - Each layer used ReLU, L2 regularization
    - Output used identity function
  - Ran for 75 epochs

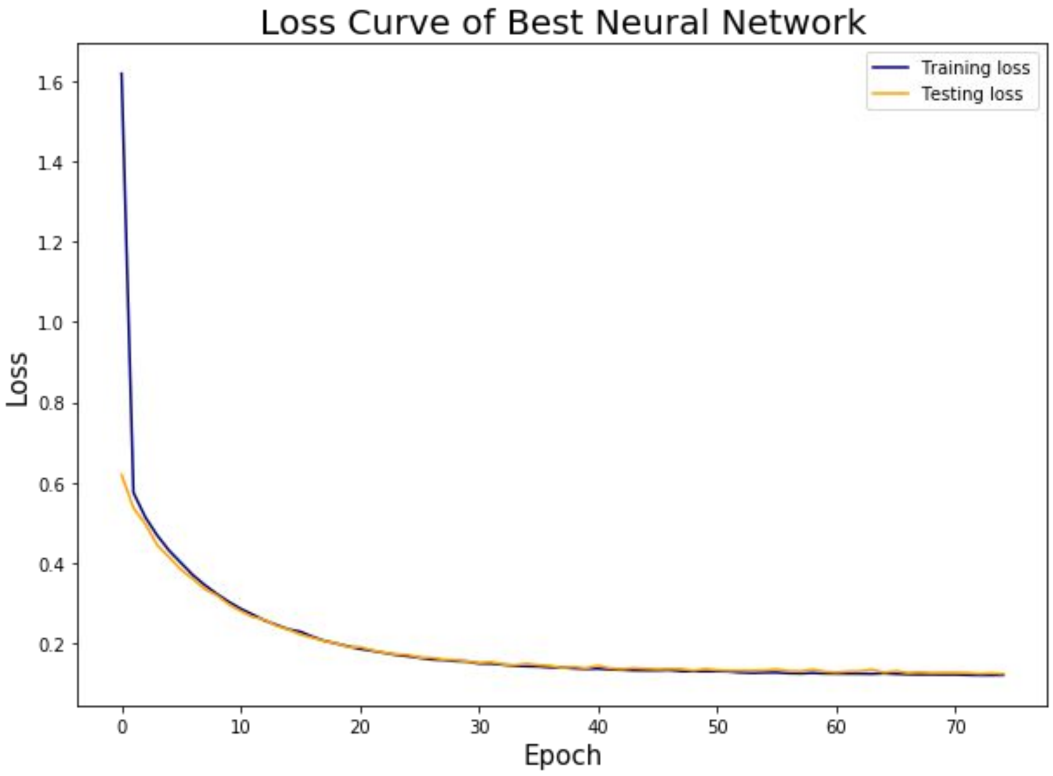
Evaluations

R^2 and Root Mean Squared Error

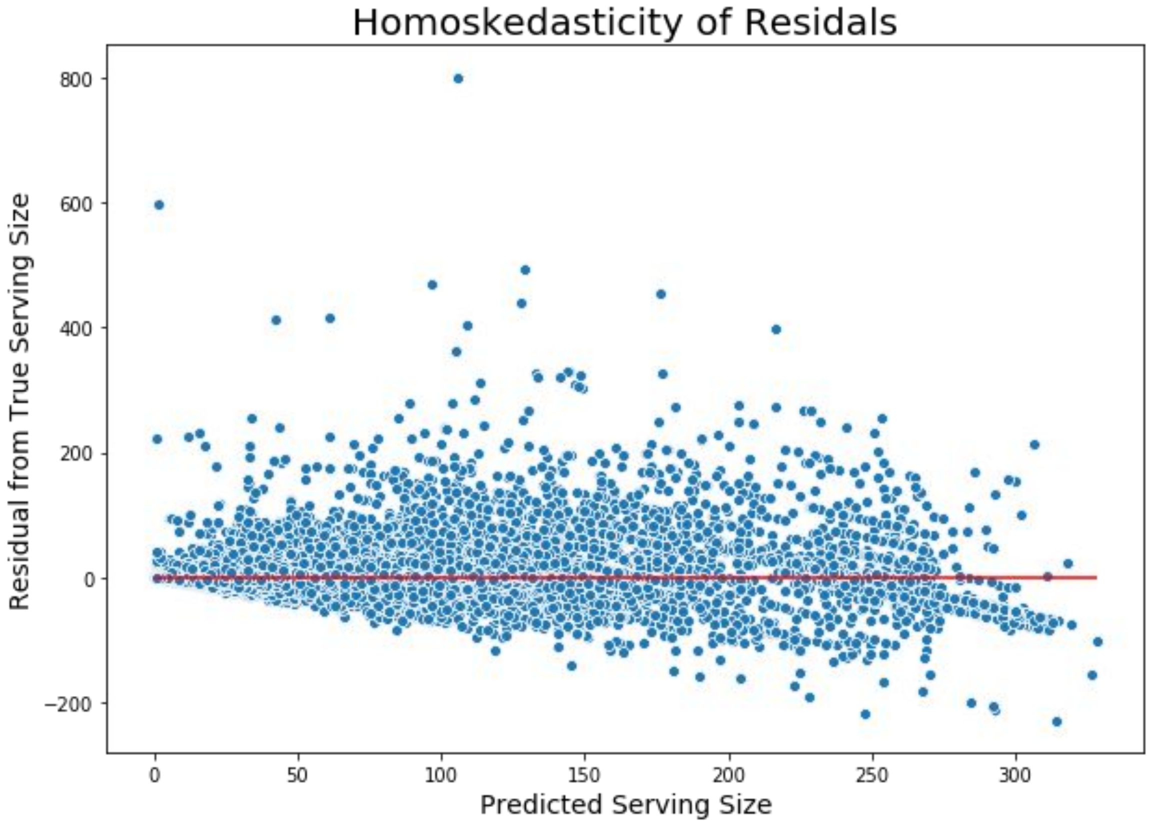
	Train R^2 %	Test R^2 %	Train RMSE	Test RMSE
OLS	69.19	68.60	31.79	32.32
Ridge	69.19	68.60	31.79	32.33
Lasso	69.19	68.58	31.79	32.33
FF Neural Net	**81.53**	**81.00**	**24.61**	**25.15**

- FFNN maximizes R^2 and minimizes RMSE
- Adjusted R^2 differed by < 0.1%

FFNN – Smooth Loss Over Epochs



Errors are Uniform and Normal



Wrap-up

Conclusions

- FFNN produced the best model
  - Leaves no interpretation of details
- Model has tendency to underpredict
  - May be affected by inedible portions
- Many other factors could have influence
  - Packaging, target market, sale price

Recommendations

- Further exploration of FFNN parameters
- Better utilize NLP for ingredients
- Attempt other model types
  - Decision trees, SVM
- Maintain dataset through USDA updates

Thank You!

Questions?

Sources

- https://data.nal.usda.gov/dataset/usda-branded-food-products-database
- https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=101.9
- https://dietarysupplementdatabase.usda.nih.gov/ingredient_calculator/help.php