Exploring Open Food Facts

Inna Munroe, July 17 2018

What is **Open Food Facts?**

- It is an open international database of products
- They also have data available in a csv format on: Kaggle.com
- Information about ingredients, origins, brands, retailers, categories and nutritional facts

Two Problems

1. **Regression problem**: can we predict energy content of a product if we only know its ingredients and serving size?

2. **Classification:** can we correctly identify food categories based on nutritional information and serving size?

Energy is measured **in kJ**

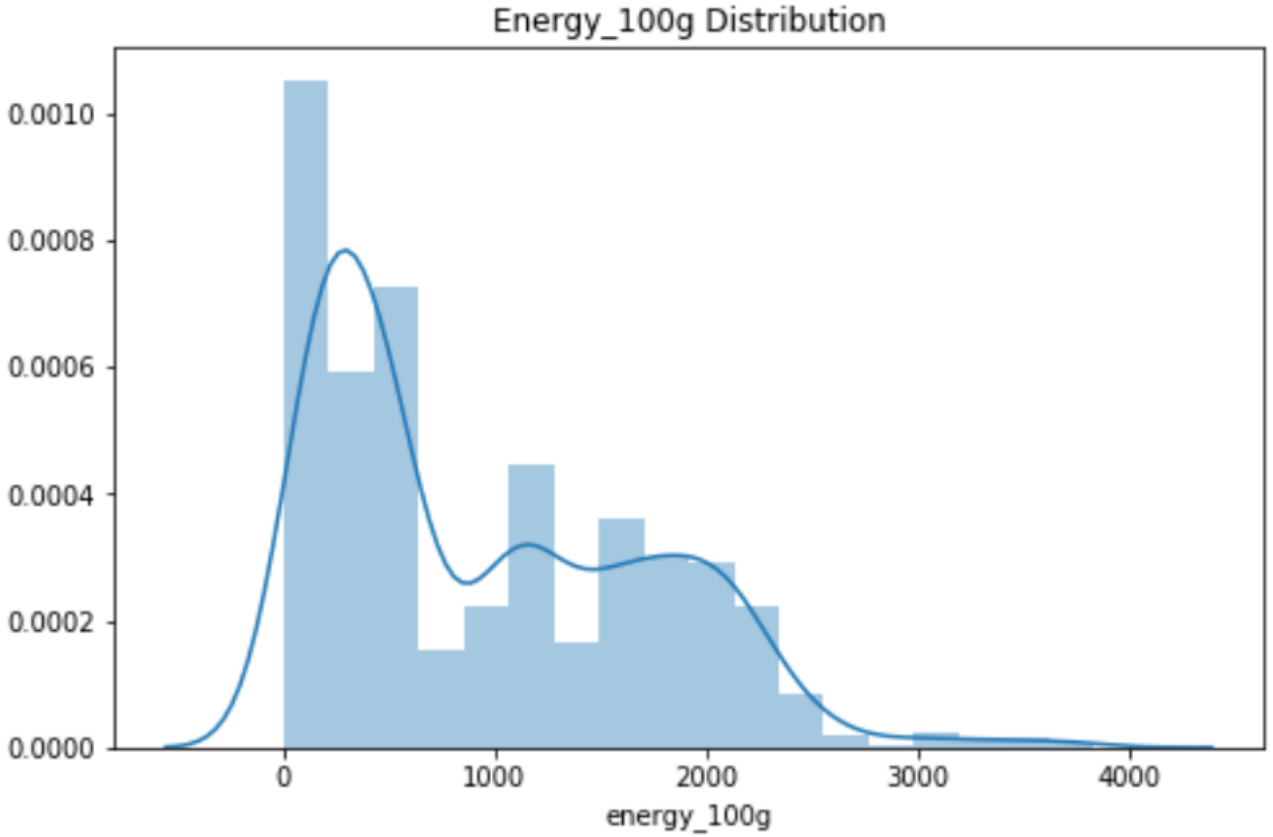
Metrics

1 Calorie = 4.184 kJ

Regression

- Two countries dominated the dataset: France and the US

- Eventually, the subset from the **US** was selected to narrow down the number of languages used for the ingredients



The target variable is not normally distributed

Non-linear transformations were attempted in the trial stages and not implemented in the final models

They did not lead to improvements in the best models

Energy per 100 g

Features

- **Serving size**: complications included high variations in servings (cups, grams, mls, table spoons etc.
- Serving size was processed through RegEx (a lot of variation in spelling and spaces)
- **Ingredients text:** text was parsed, binary features were created based on the vocabulary items that appear in the training set
- **Food categories:** dummies for categorical variables

Main Metrics Used

- RMSE (root mean square error)

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

- It tells us what does a typical error of our model looks like
- Sensitive to large errors, useful when large errors are particularly undesirable

Models Attempted

• **Linear Models**:	**Training Set Cross Validation**	**Test Set**
Ridge Regression:	408.3 kJ	388.21
Lasso Regression:	448.62 kJ	428.78
Elastic Net:	405.91 kJ	387.73

Models Attempted

- **Tree Models**:

	Training Set Cross Validation	Test Set
Random Forest:	334.76 kJ	320.92 kJ
XGBoost Regression:	326.05 jK	330.70 kJ

Mean Absolute Error and R^2

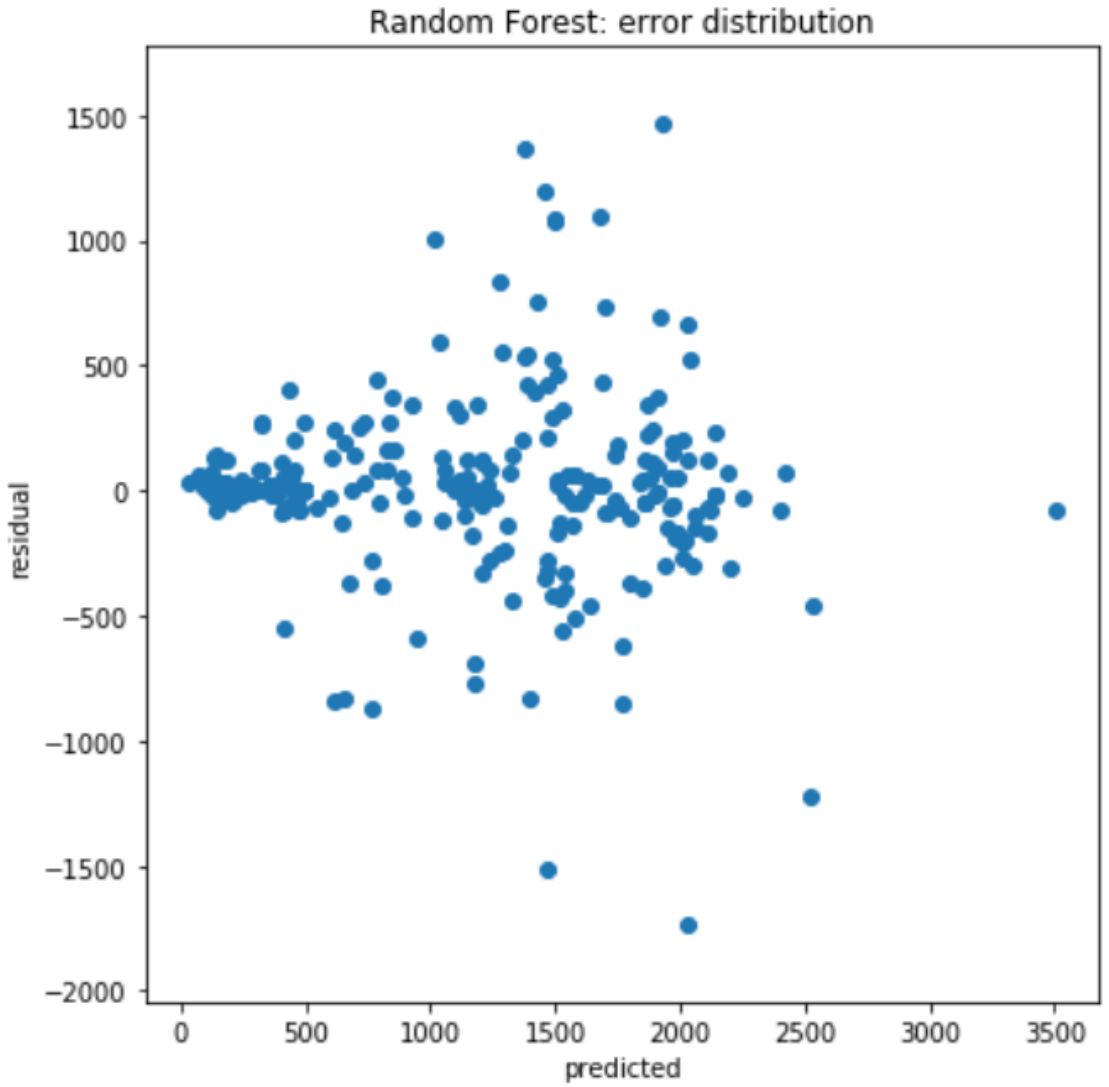
$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n} |y_j - \hat{y}_j|$$

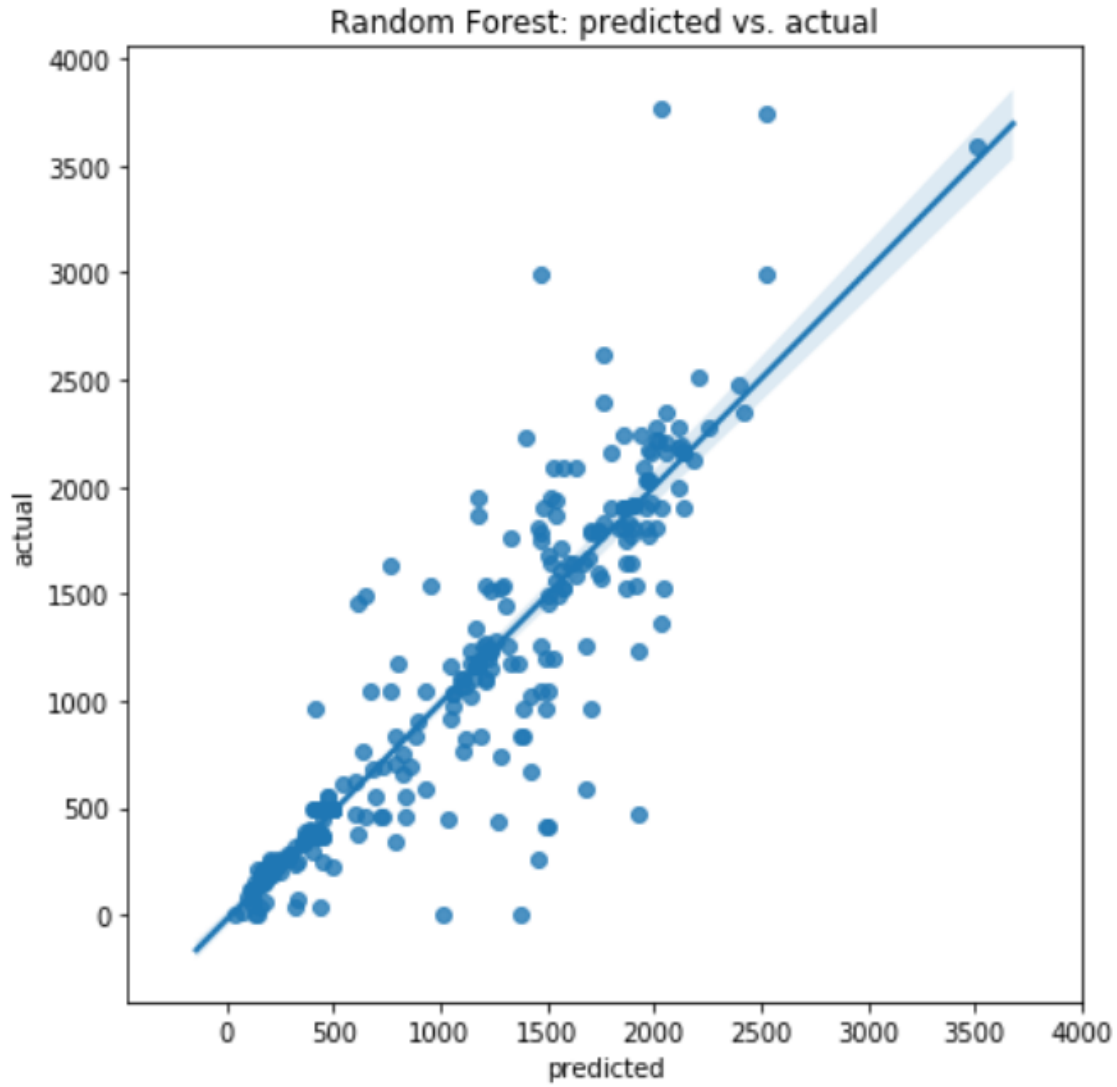
XGBoost:	179.85 kJ	0.8149
Random Forest:	169.22 kJ	0.8257

Best Result

- RMSE: 320.92 kJ (76 Calories) per 100 g
- MAE: 169.22 kJ (or 40.4 Calories) per 100 g

Errors Analysis





Improvements

- Better Text Parsing (time limit to handle all the punctuation complexities)
- More model precise parameter finetuning
- Information about the weight of each ingredient might be needed for more predictive power

Potential Uses

- Voice assistants (interactive Calorie prediction based on ingredients)
- Diet Apps (diet optimization)

Classifier

- Currently, the dataset has 270624 uncategorized food items

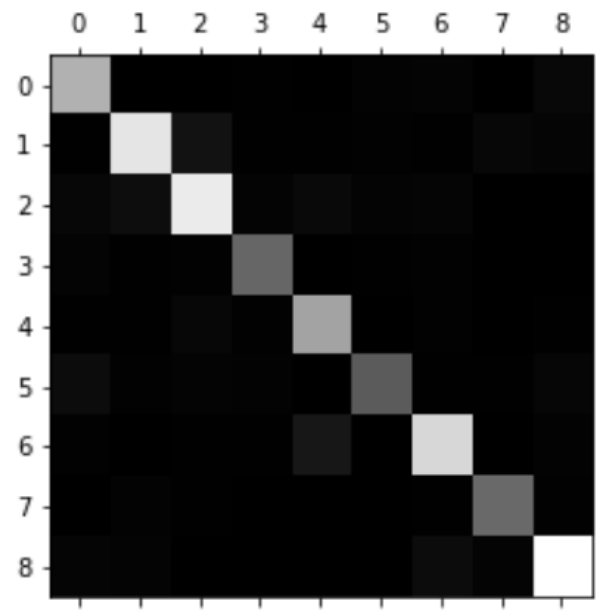
- We could use machine learning to classify items based on their nutritional content and serving size

unknown	270624
Sugary snacks	15369
Beverages	13476
Milk and dairy products	10733
Cereals and potatoes	10097
Fish Meat Eggs	9473
Composite foods	7972
Fruits and vegetables	7861
Fat and sauces	7122
Salty snacks	3300

Tree Models Accuracy

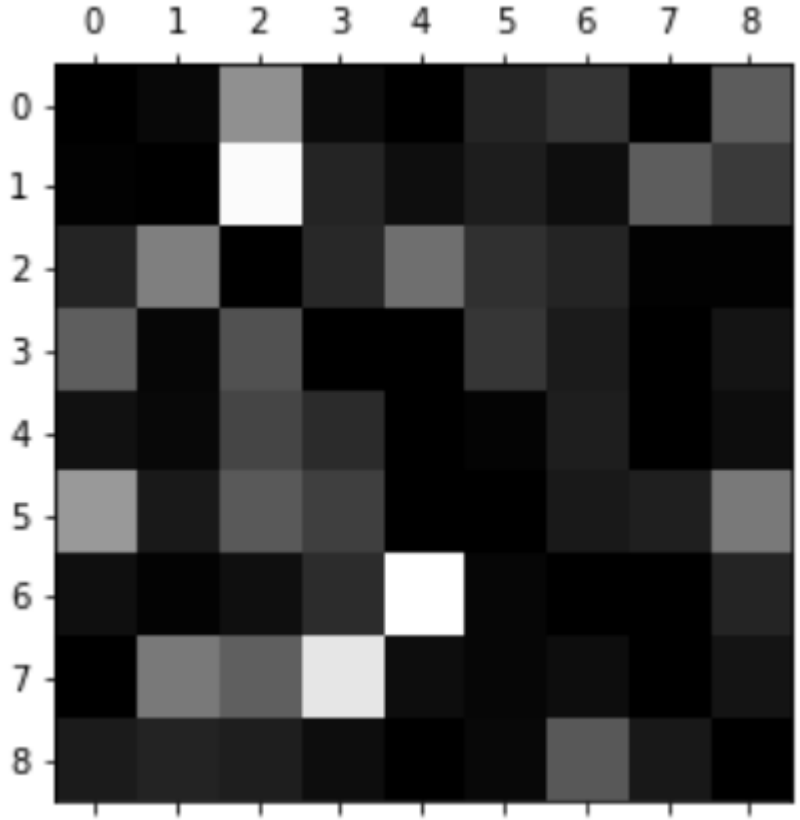
Tree Models:	**Training Set Cross Validation**	**Test Set**
Random Forest:	0.9027	0.87
XGBoost Classifier:	0.903	0.84

Random Forest Classification and Errors




```
: forest_best.classes_
```

```
: array(['Beverages', 'Cereals and potatoes', 'Composite foods',
         'Fat and sauces', 'Fish Meat Eggs', 'Fruits and vegetables',
         'Milk and dairy products', 'Salty snacks', 'Sugary snacks'],
        dtype=object)
```



Improvements

- Initial categorization might not be precise:

  e.g. fish meat eggs and milk and dairy products were frequently confused by the model

- Cereals and potatoes are confused with composite foods

Potential Uses

This kind of classifier can be used to improve the current *Open Food Facts* dataset categorization of the unknown, especially with no pictures

Beyond this dataset, it can be employed for food recommendation and automated food classification

Photography Credits

All photographs used in the PowerPoint were taken by **Marc Bell**