Predicting the number of people have tried the recipe based on the food data

BA 865

Group members

Zixing Li, Mengxin Li, Hanyu Chen, Yapei Xiong

BOSTON UNIVERSITY

Boston University Questrom School of Business

PROBLEM STATEMENT

Business goals:

- Understanding how different features can affect the result
- Building appropriate models for our data using tensorflow and keras
- Achieving website click-through rate increasing
- Generating more users to use the recipe website in the future

Project description:

- The main aim of our project is to predict the number of people who have tried the recipe based on the food recipes data using deep learning technology

DATASET

- The ‘BBC Food Recipes dataset’ was downloaded from Crawl Feeds, originally extracted from the BBC good food, and there is more than 14000+ recipes
- Consisted of a json file with **18 columns** and **14.2k rows**, and json file containing category names
- All the features are object data type, and 6 columns has more than 1k null values

Column	Non-value	Dtype
total_time	2011	object
cook_time	3859	object
serving	2076	object
crawled_at	0	object
description	0	object
title	0	object
url	0	object
nutritions_info	2197	object
image	0	object
ingredients	137	object
uniq_id	0	object
source	0	object
author	0	object
prep_time	3090	object
published_date	0	object
keywords	360	object
total_ratings	53	object
instructions	1959	object

Source: https://crawlfeeds.com/datasets/bbc-food-recipes-dataset

DATA CLEANING

- Drop columns: too much to handle or related to other column

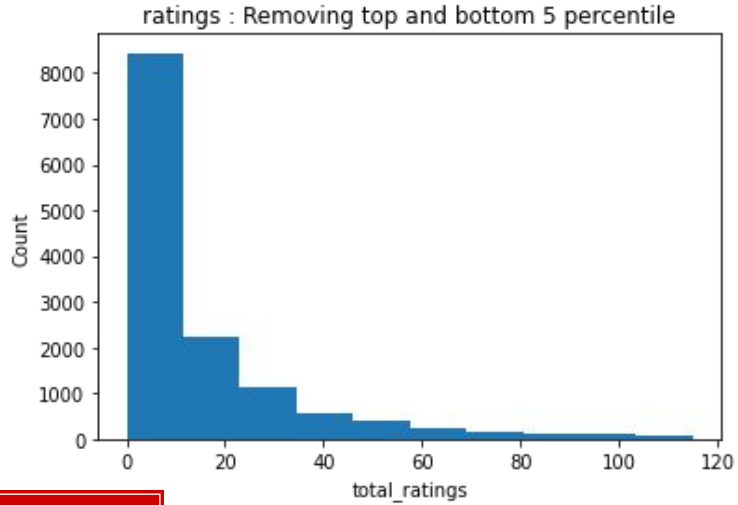
‘uniq_id’	‘prepare_time’
‘cook_time’	‘instructions’
‘crawled_at’	‘url’
‘source’	

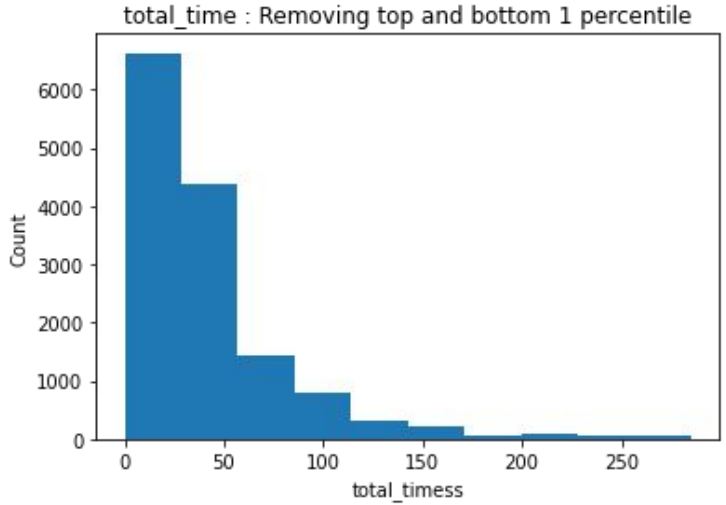
- Deal with all the null-values, using timestamp to convert the unusual time type to minutes
- Using LabelEncoder to encode the ‘author’
- Using TfidVectorizer to token the text columns: ‘description’, ‘keywords’, ‘ingredients’

- Split the ‘nutrition_information_values’ column into 8 columns: 'calories', 'fat', 'saturatedFat', 'carbohydrate', 'sugar', 'fiber', 'protein', 'sodium', and deal with the null value
- Drop the columns that we tokenized

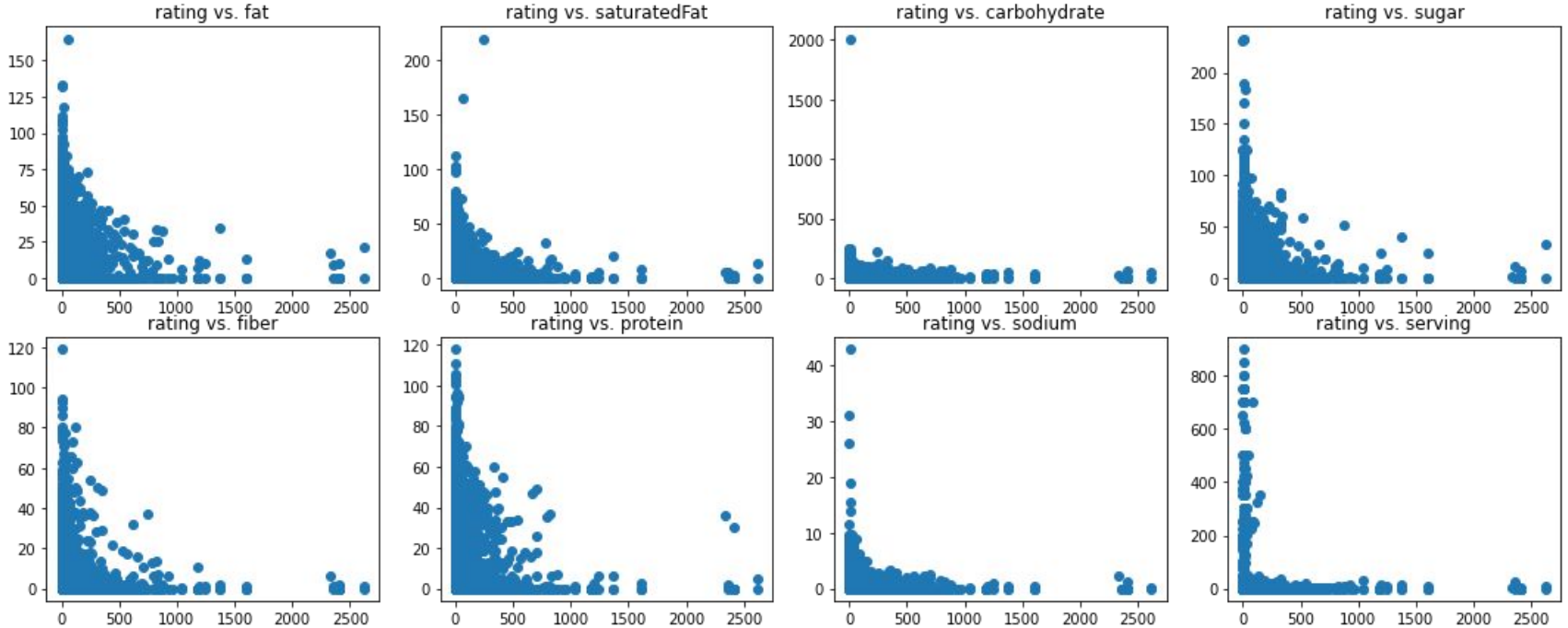
EDA

- To see the true distributions of our target label and important feature, we remove the top and bottom 5 and 1 percentile.

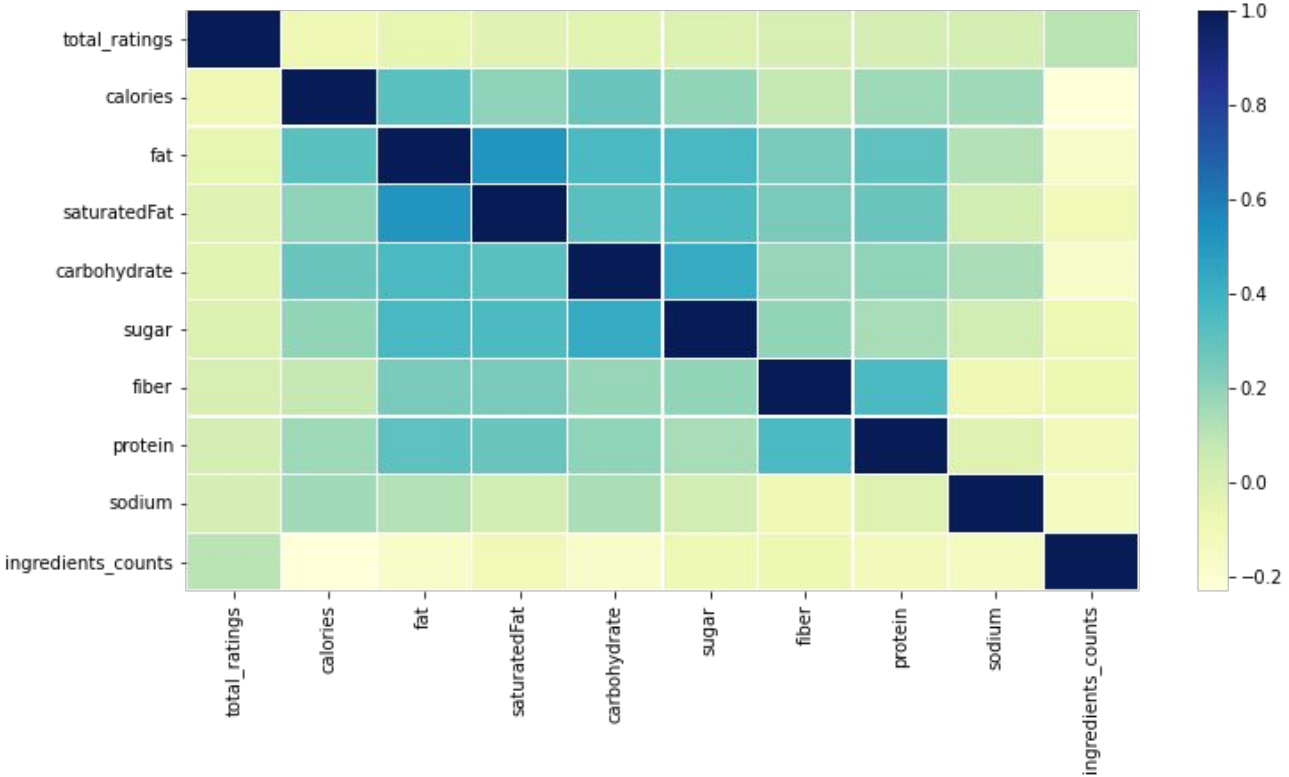




EDA


rating vs. fat
rating vs. saturatedFat
rating vs. carbohydrate
rating vs. sugar
rating vs. fiber
rating vs. protein
rating vs. sodium
rating vs. serving

EDA


total_ratings
calories
fat
saturatedFat
carbohydrate
sugar
fiber
protein
sodium
ingredients_counts
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

DEEP LEARNING - IMAGE

Image Pre-processing

- Dealing the problem that some image urls not able to access.
- Delete the rows that contains dead links

```
[24] deadlist = ['https://images.immediate.co.uk/production/volatile/sites/30/2021/05/Five-bean-chilli--e4da69b.jpg',
      'https://images.immediate.co.uk/production/volatile/sites/30/2020/09/Snacks-Chickpeas_00154-31614c9.jpg',
      'https://images.immediate.co.uk/production/volatile/sites/30/2013/05/vegan-jambalaya-01e5bde.jpg',
      'https://images.immediate.co.uk/production/volatile/sites/30/2021/08/easy-jam-sponge-aff767d.jpg',
      'https://images.immediate.co.uk/production/volatile/sites/30/2020/08/passion-fruit-cake-896b246.png',
      'https://images.immediate.co.uk/production/volatile/sites/30/2020/10/Chicken-stew-425b8e0.jpg',
      'https://images.immediate.co.uk/production/volatile/sites/30/2017/06/Tomato-soup-5cf0912.jpg',
      'https://images.immediate.co.uk/production/volatile/sites/30/2021/07/Griddled-courgettes-4c9269d.jpg',
      'https://images.immediate.co.uk/production/volatile/sites/30/2021/08/vegan-tiramisu-1dab5a8.jpg',
      'https://images.immediate.co.uk/production/volatile/sites/30/2020/11/Butter-chicken-8f9fd05.jpg',
```

- Resize the image size to (200,200,3) and convert to numpy arrays
- Replicate grayscale image 3 times to make it like a RGB photo
- Split our dataset with 2100 for training and 900 for testing

DEEP LEARNING -NUMERIC

Numeric Pre-processing

- MinMax scaler numeric values

```
["total_time","calories","fat",'saturatedFat','carbohydrate','
sugar','fiber','protein','sodium']
```

- Labels: total_ratings; Predictors: Total_time, serving, published_date…..
- train/test split (70%/30%)
- One hot encoded as mentioned in previous

DEEP LEARNING - CNN

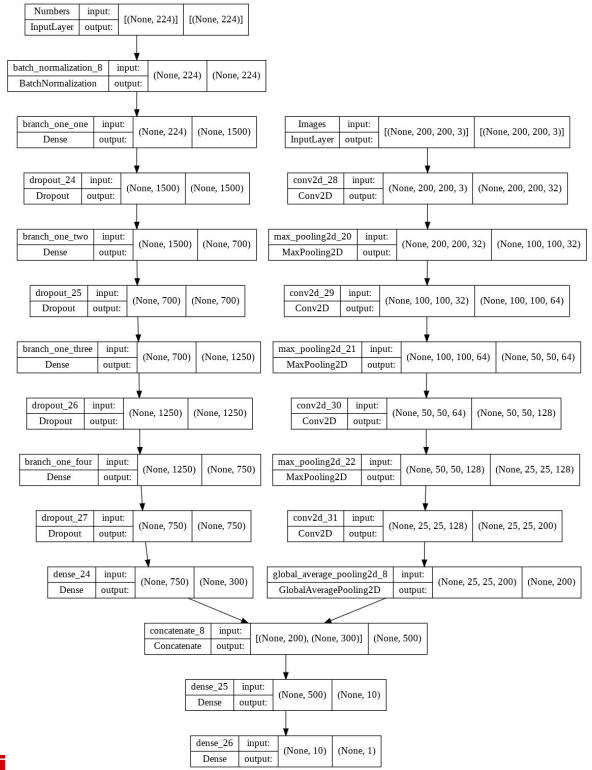
Model Building - Overfitting

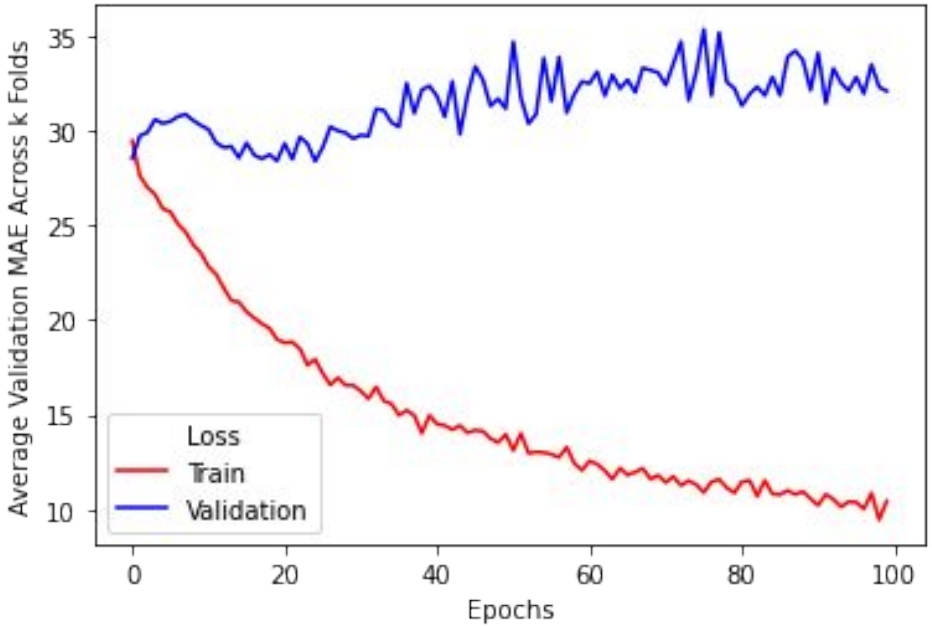
- Multiple layers
- Droup-out layers for numeric layers (0.5)
- Enlarging kernel sizes
- Applying Kernal regularizer and bias regularizer on both convolutional layer and numeri layers
- ‘Relu’ activation function. (Not a categorical question)

Model fitting - 4 fold cross validation

- Epoch: 100
- Batch size: 75

MODEL





BOSTON UNIVERSITY

CONCLUSION

Limitations:

- MAE train lower than 10 while validation loss remained high
- Overfitting issue still remains, potential reasons could be cutting a lot of data points.
- Data pre-processing takes up a lot of time and computer glitched brought technical issues.

Bright sight:

- New idea in combination of text analytics, different kinds of ingredients and pictures of the food
- Increase the accuracy of the model by bringing in more data points
- Help those choose recipes wisely and those who have an intention to publish blog. :)

Thank You for Watching!