Tasty Bites: Machine Learning Solution

Liam Haney

Project Goals

Request:

- Predict which recipes will lead to high traffic
- Correctly predict high traffic recipes 80% of the time

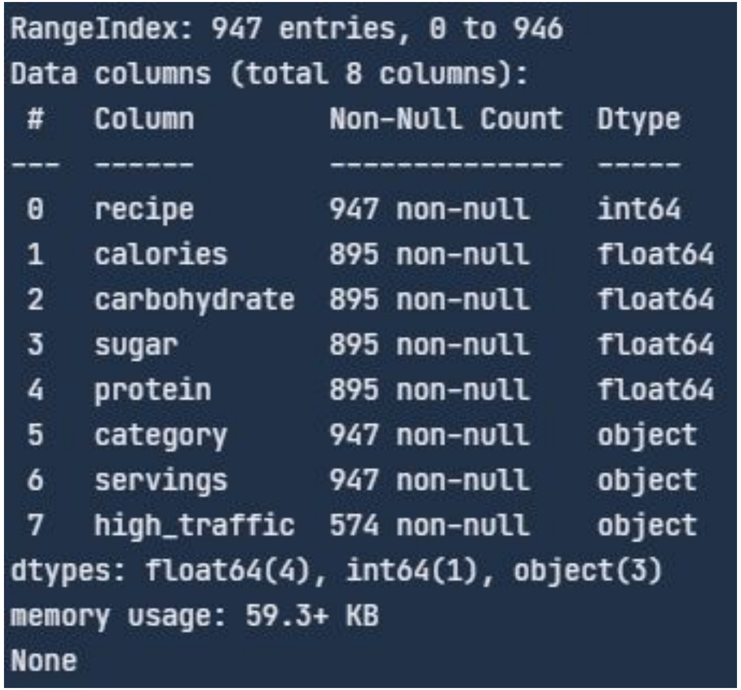
Solution:

- Create predictive model using provided data

Provided Data

947 entries with information across 8 columns

- 1 unique identifier column
- 1 target variable column
- 6 feature columns
- Some entries missing data

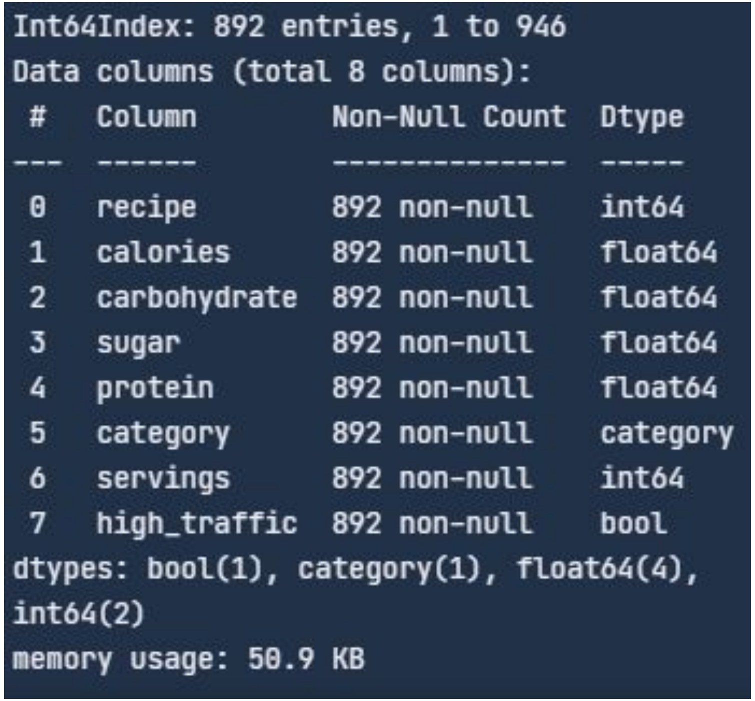


```
RangeIndex: 947 entries, 0 to 946
Data columns (total 8 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   recipe        947 non-null    int64
 1   calories      895 non-null    float64
 2   carbohydrate  895 non-null    float64
 3   sugar         895 non-null    float64
 4   protein       895 non-null    float64
 5   category      947 non-null    object
 6   servings      947 non-null    object
 7   high_traffic  574 non-null    object
dtypes: float64(4), int64(1), object(3)
memory usage: 59.3+ KB
None
```

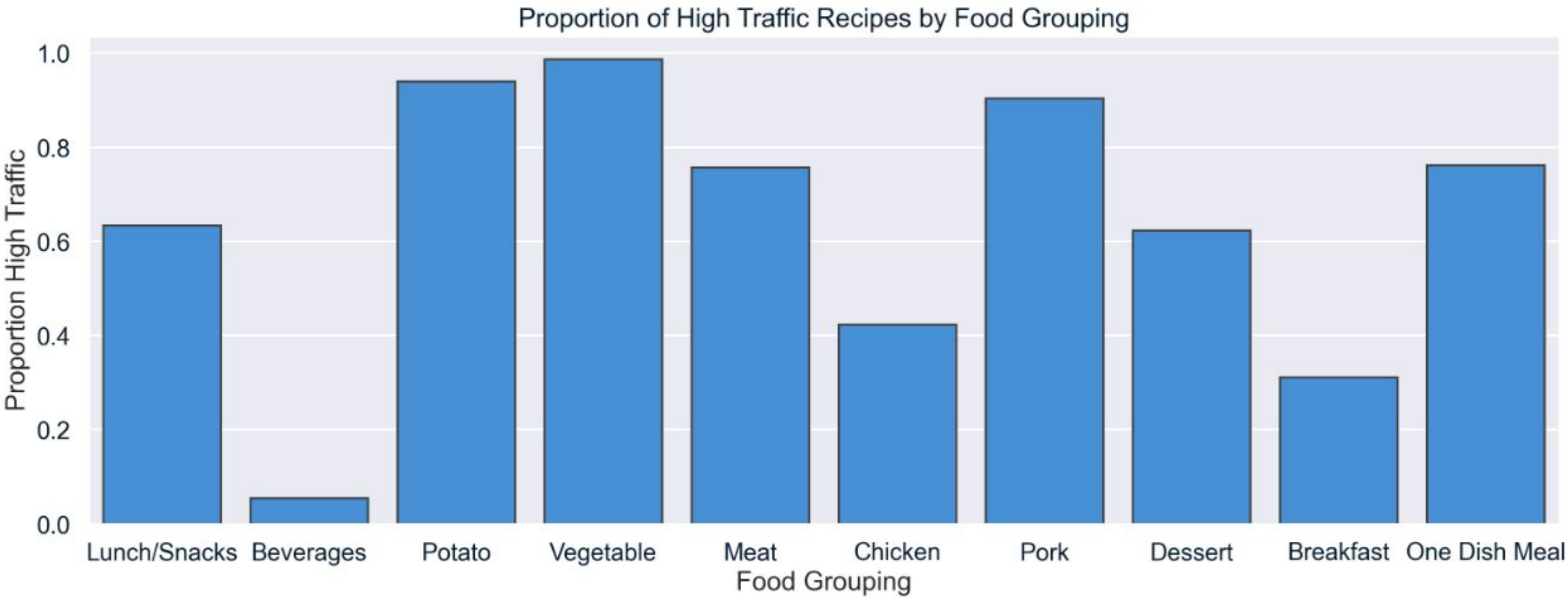
Data Cleaning

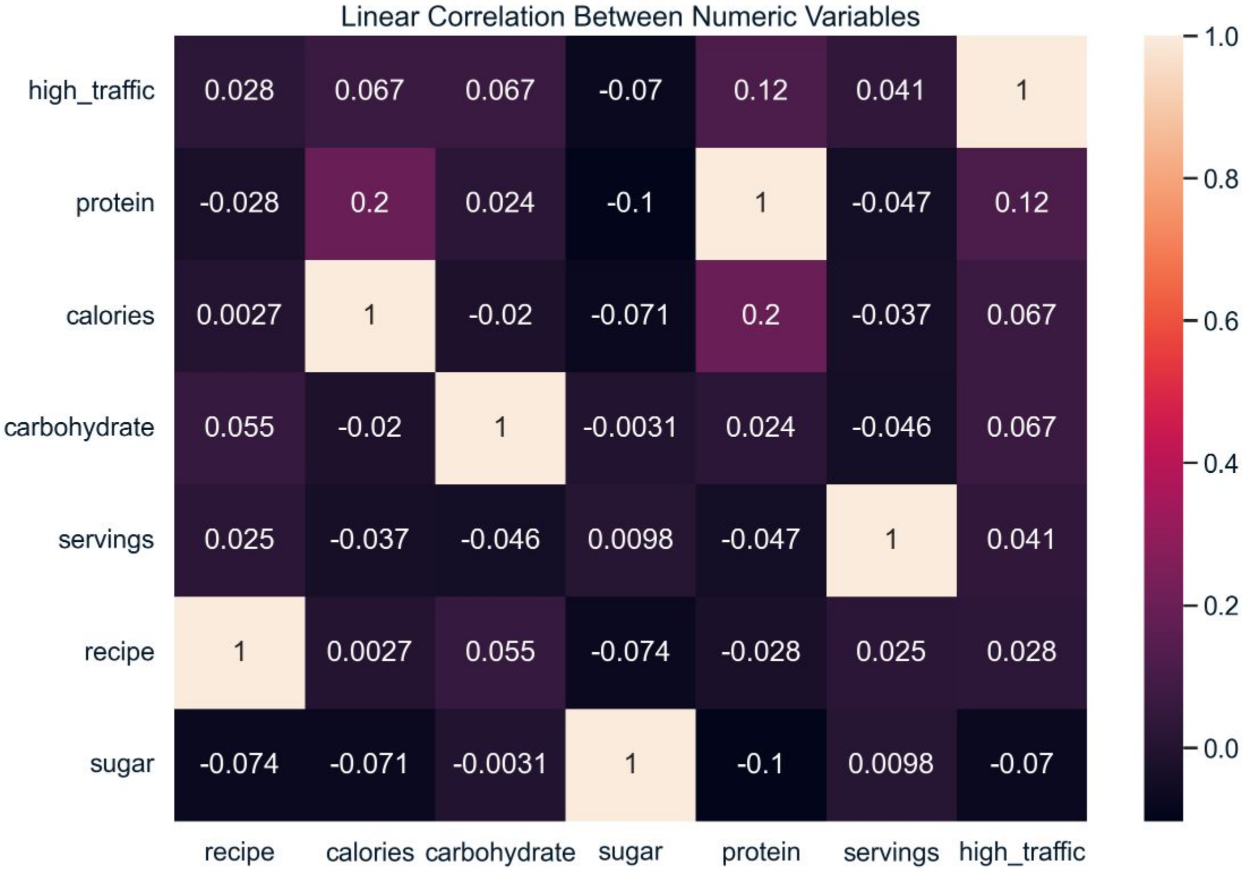
892 entries across 8 columns

- Removed 55 entries
  - 52 missing numeric data in 4 columns
  - 3 ambiguous 'servings' ("as a snack")
- Consolidated extra food grouping, 'Chicken Breast', with 'Chicken'
- Converted to correct data types e.g. 'high_traffic' is now True or False



```
Int64Index: 892 entries, 1 to 946
Data columns (total 8 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   recipe       892 non-null    int64
 1   calories     892 non-null    float64
 2   carbohydrate 892 non-null    float64
 3   sugar        892 non-null    float64
 4   protein      892 non-null    float64
 5   category     892 non-null    category
 6   servings     892 non-null    int64
 7   high_traffic 892 non-null    bool
dtypes: bool(1), category(1), float64(4),
int64(2)
memory usage: 50.9 KB
```


Proportion of High Traffic Recipes by Food Grouping
Proportion High Traffic
1.0
0.8
0.6
0.4
0.2
0.0
Lunch/Snacks
Beverages
Potato
Vegetable
Meat
Chicken
Pork
Dessert
Breakfast
One Dish Meal
Food Grouping



Linear Correlation Between Numeric Variables

	recipe	calories	carbohydrate	sugar	protein	servings	high_traffic
high_traffic	0.028	0.067	0.067	-0.07	0.12	0.041	1
protein	-0.028	0.2	0.024	-0.1	1	-0.047	0.12
calories	0.0027	1	-0.02	-0.071	0.2	-0.037	0.067
carbohydrate	0.055	-0.02	1	-0.0031	0.024	-0.046	0.067
servings	0.025	-0.037	-0.046	0.0098	-0.047	1	0.041
recipe	1	0.0027	0.055	-0.074	-0.028	0.025	0.028
sugar	-0.074	-0.071	-0.0031	1	-0.1	0.0098	-0.07

Similarly insignificant results for logarithmic correlation

Model Development

To reach goal

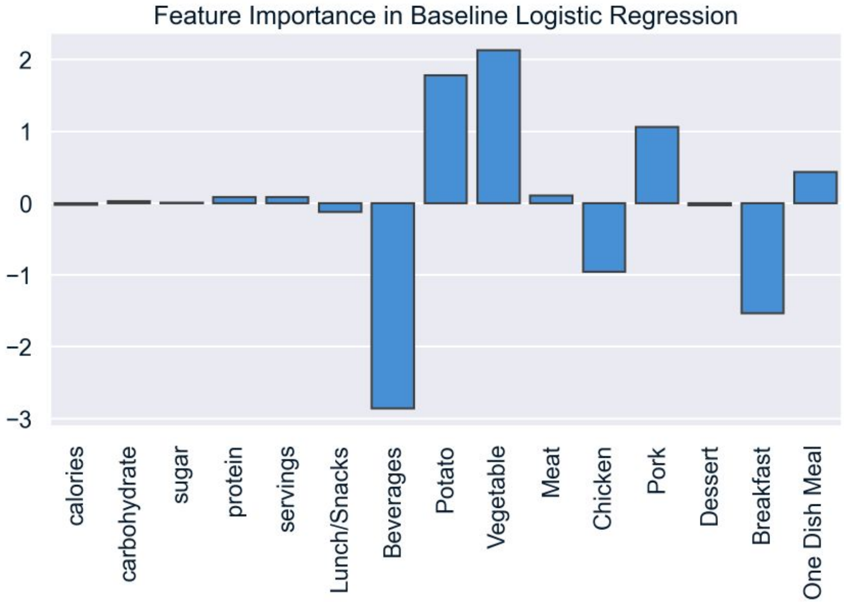
- Binary classification model that correctly predicts 80% of high traffic recipes

Chosen models

- Logistic Regression Model (Baseline)
- Random Forest Classifier

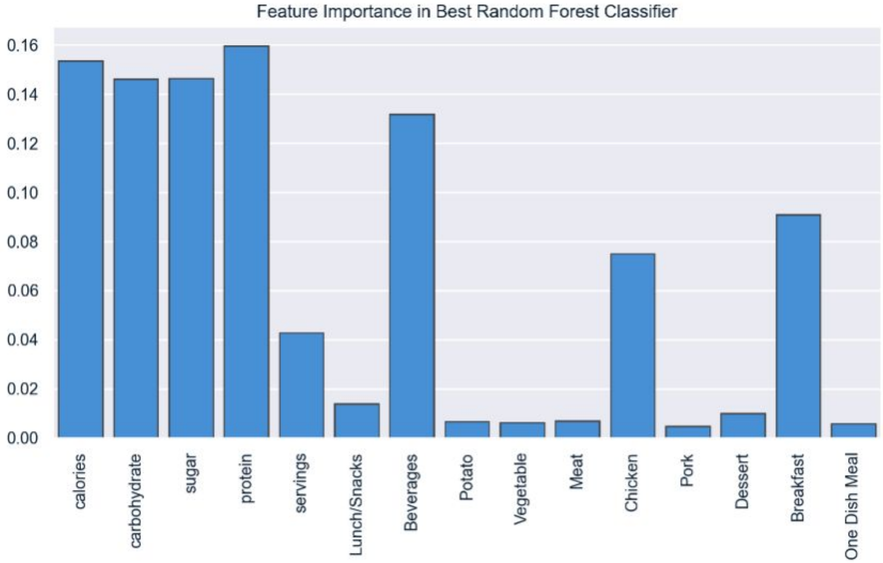
Logistic Regression Model

Less complex, readily captures simple relationships



Random Forest Classifier

More complex and adaptable



Model Evaluation

Evaluation

- Accuracy: baseline correctness, (note class imbalance >60% True)
  - 0 - 100%
- Recall: percent of high traffic recipes that were correctly identified
  - 0 - 100%
- ROC AUC: True vs. False Positive rate at various classification thresholds
  - 0.5 - 1.0
  - Random guessing to perfect predictions

Model Comparison

Metric	Logistic Regression	Random Forest
Accuracy	74.6%	**75.4%**
Recall	81.3%	**81.9%**
ROC AUC	0.730	**0.738**

Business Application

Key Performance Indicator (KPI)

- Recall of 81.9% exceeds requested 80%
- Implement model, compare results of model predictions vs. manual recipe selection
- Dashboard application

Current model can be improved with some clarification and additional feature data

Moving Forward

Additional Data Collection

- Missing data: 52 entries across 4 columns, can be corrected?
- 'Time to make', 'Cost per serving', 'Ingredients', perhaps relevant?

Generate New Features

- 'category': kept as 10 provided food groupings, values not exclusive and should probably be split
- 'high_traffic': alter threshold to identify only most important features/change to regression problem

Thank you

Keep in touch!
Phone: 123-123-1234
Email: email@email.com