Report by Francisco Baptista

Recipes Traffic

TASTY BYTES

The structure of this presentation

01 The task and objective

02 Summary of the work

03 The metrics used and why

04 Recommendations

TASK

Create a model to help predict which recipes will lead to high traffic because more traffic means more subscriptions.

OBJECTIVE

Correctly predict high traffic recipes 80% of the time.

Some takeaways from the data

01 There were only missing values (52) on the nutritional information.

02 The overall size of the dataset was quite small: 947 recipes.

03 The relevant data is compost by two categorical columns (servings and category) and 4 numeric (calories, sugar, protein and carbohydrate).

04 The data contains, close to, 60% of high traffic recipes (535).

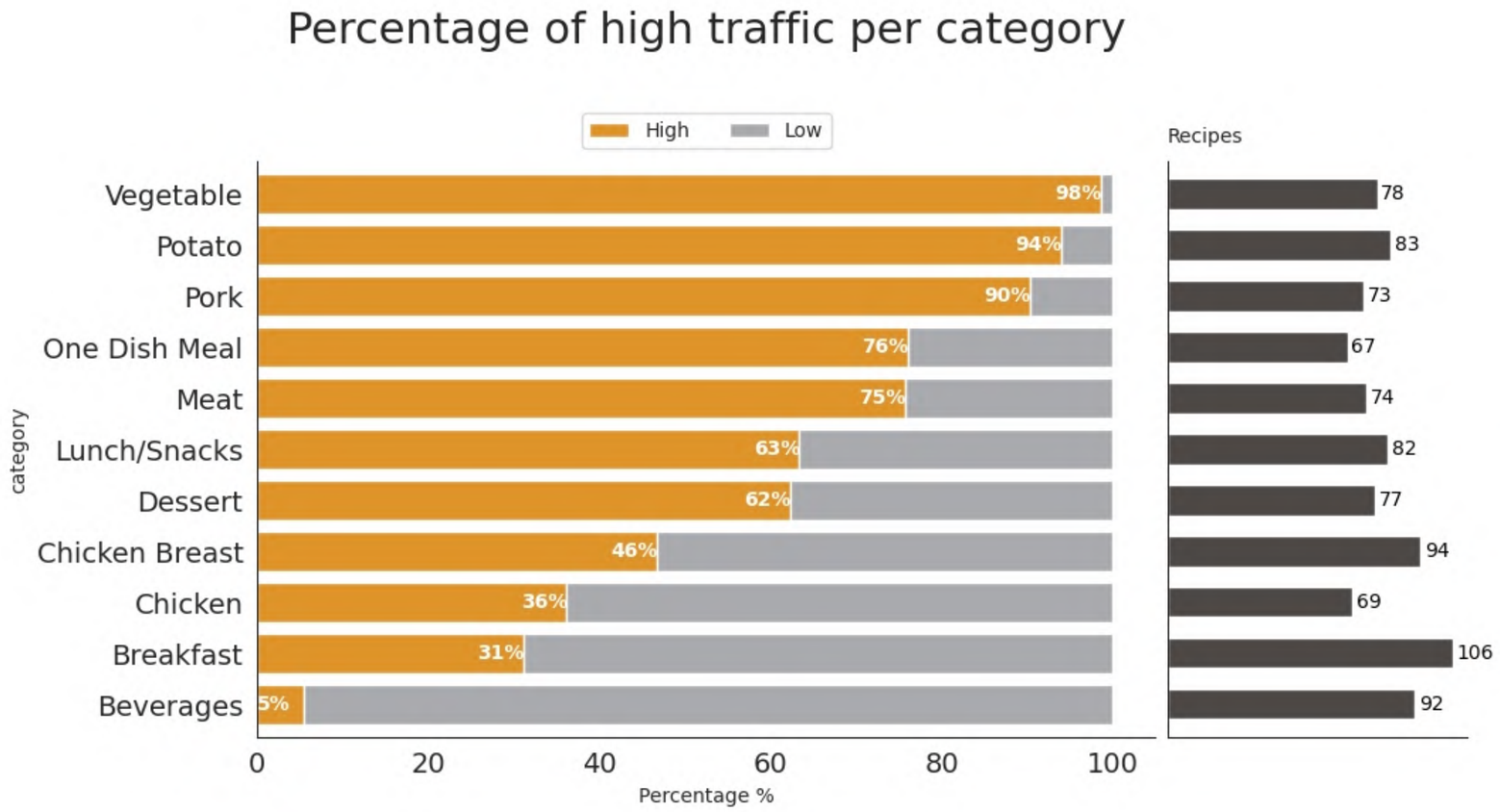
The data analysis shows

The recipes category has the strongest connection with the traffic as we can on the chart below

Recipes with vegetables, potatoes and pork have 90% or more high traffic.

On the other end, Beverages have the least recipes with higher traffic.

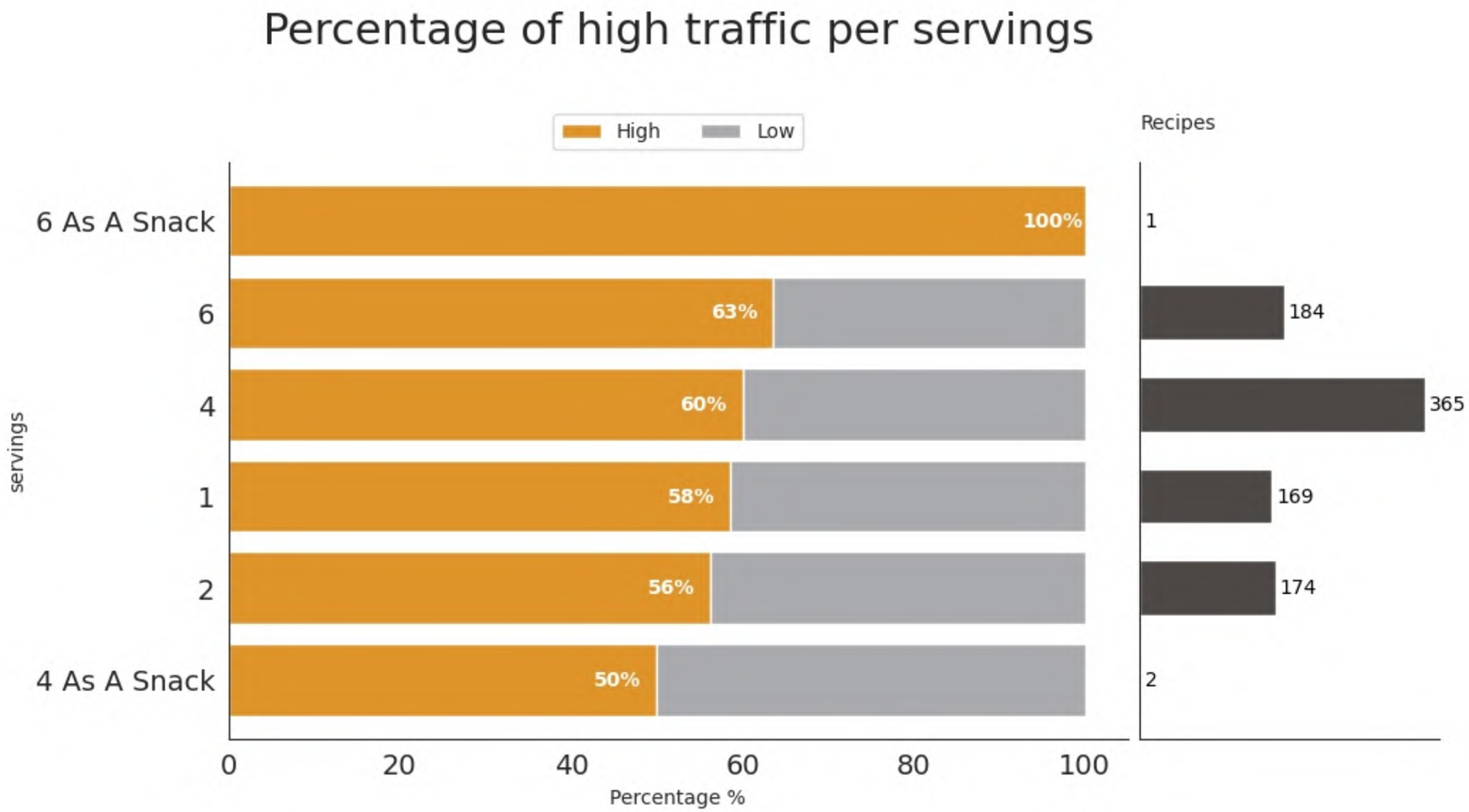
I included the number of recipes for each category, so we could have an idea of the percentage weight. The next chart shows how this can be important.



There's no number of servings with a significant impact on traffic.

Recipes with 6 servings have more traffic but overall, every serving size has 50% or more, high traffic recipes.

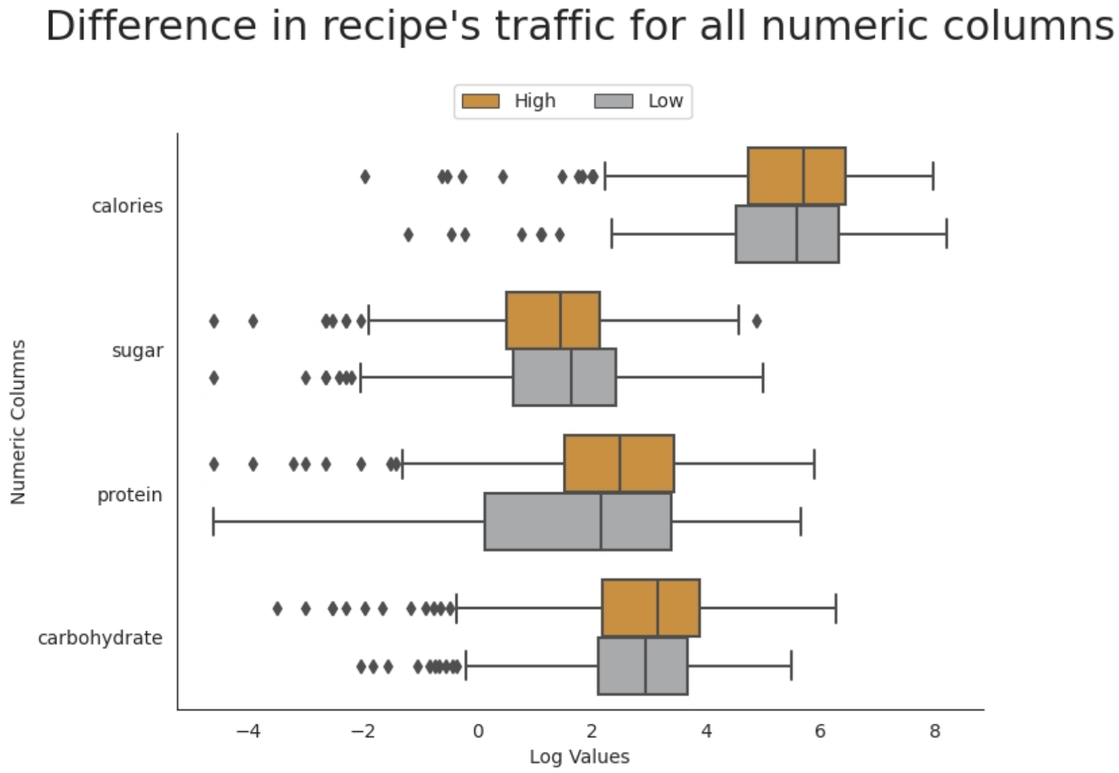
The number of recipes in this chart really helps us put in to perspective the weigh of the servings "as a snack" because the quantity in so small it introduces a lot of volatility to the percentages.



All the numeric columns have similar distributions and have little to no correlation with a recipe's traffic.

For that, we can already assume that, those will not help our model predict if a recipe has higher traffic.

One explanation could be because, in general, people tend to not pay as much attention to the nutritional information, unless they have a strict diet of some sort.



Model Development

Summary

1

Since the objective was to develop a model to predict between two outcomes (high and low traffic) I evaluate 4 classification models being Logistic Regression the best performing one.

2

Several data manipulations were tried but nothing improve the model, so the data was feed to the model as it was.

3

As was shown on the data analysis previously, the nutritional information did not contribute significantly to the model, so the final model only uses the categorical features of the data.

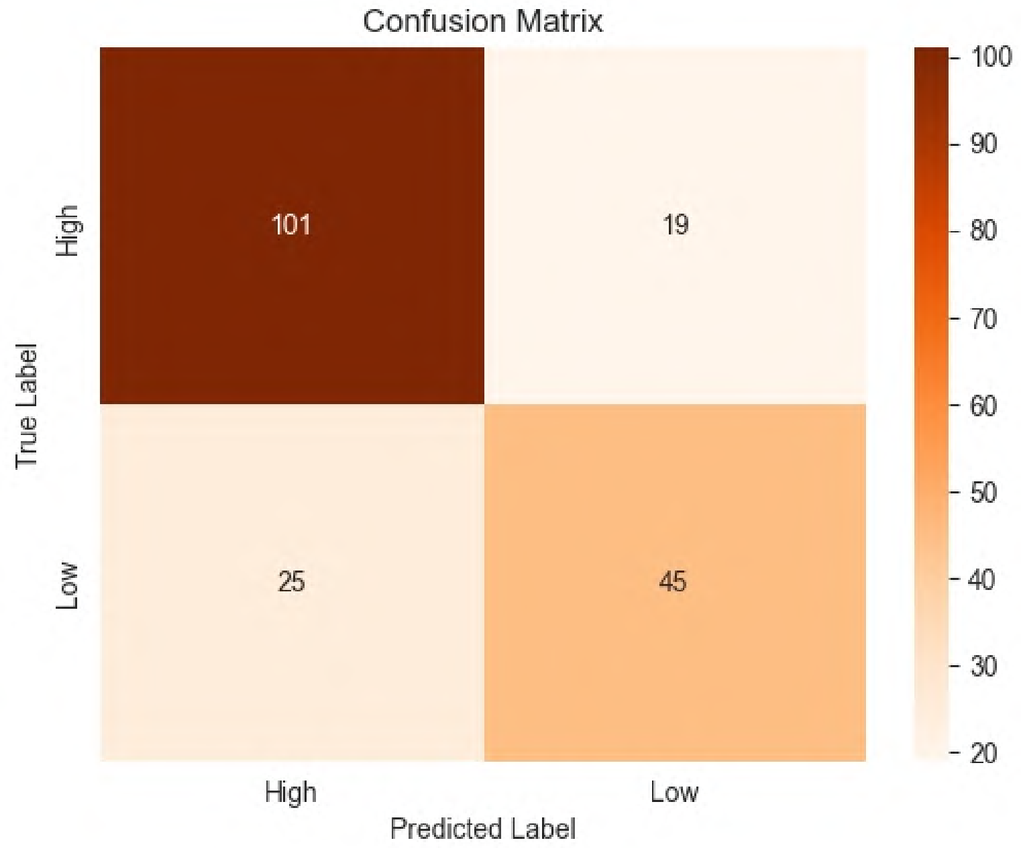
4

The metric I found most appropriate to this task was precision, given that the request was to accurately predict 80% of high traffic recipes, while considering other metrics, such as accuracy, for a more comprehensive understanding of overall performance, since my aim was to develop the most robust model possible.

Based on this chart we can conclude that the model performed as requested, having:

Predicted correctly 101 out of 120 high traffic recipes.

This yields a precision score of 84.17%, surpassing the requested benchmark by 4.17 percentage points.



Tomato Soup

Servings: 4
Time to make: 2 hours
Category: Lunch/Snack
Cost per serving: $

Nutritional Information (per serving)	
Calories	123
Carbohydrate	13g
Sugar	1g
Protein	4g

Ingredients:

- Tomatoes
- Onion
- Carrot
- Vegetable Stock

Method:

1. Cut the tomatoes into quarters....

Recommendations

Even though the model scored +80% on the high traffic recipes it doesn't seem highly robust for the future since it's highly dependent on one column ("Category").

I recommend start using more data from each recipe, such as:

- Time to make: People may choose recipes based on their preference for faster or longer cooking times;
- Cost per serving: Not knowing the financial condition average of the users, something that's pretty common nowadays and in times of financial tightening is for cheaper recipes to have higher traffic;
- Ingredients: In my view, this is a great filter to have on our data, because enables us to conduct a more in-depth analysis and categorize recipes based on specific ingredients. This is especially important as the category column lacks detail.

If you have any questions please let me know.

Thank you