RECIPE SITE TRAFFIC

How to correctly predict a high traffic recipe?

Why are we doing this?

Having our BUSINESS in mind
We want to drive traffic to our website to get subscribers.

We detect a PROBLEM
The product manager currently selects daily homepage recipes based on personal preference.

There's a NEGATIVE IMPACT
Our website traffic is highly variable.

This highlights the need for a more data-driven strategy.

Generate a model that predicts recipes with high traffic.

What do we have?

Historical Data

Product Team

Data Team

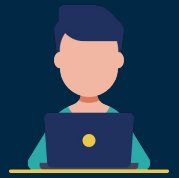
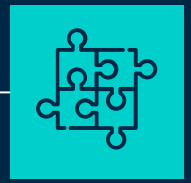
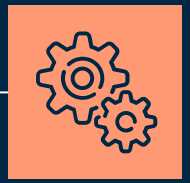


TABLE OF CONTENTS



01

DATA VALIDATION & EDA



02

MODEL DEPLOYMENT & EVALUATION

03

HOW TO HELP THE BUSINESS?

1. DATA VALIDATION

Column	Non-Null Count	Dtype
recipe	947 non-null	int64
calories	895 non-null	float64
carbohydrate	895 non-null	float64
sugar	895 non-null	float64
protein	895 non-null	float64
category	947 non-null	object
servings	947 non-null	object
high_traffic	574 non-null	object

The available data contains 8 main variables:

recipe: numeric	Unique identifier of recipe
calories: numeric	Number of calories
carbohydrate: numeric	Amount of carbohydrates (g)
sugar: numeric	Amount of sugar (g)
protein: numeric	Amount of proteins (g)
category: character	Type of recipe.
servings: numeric	Number of servings for the recipe.
high_traffic: character	Indicates if the traffic was high.

The record identifier.

Features to be considered in the model.

Target variable to predict.

1. DATA VALIDATION

Data needs to be robust in order to train a ML model. To ensure the quality of the data, three main steps have been performed.

1. Removing duplicates.
2. Checking the values of each columns and adjusting their datatypes.
3. Assessing null values.

Column	Non-Null Count	Dtype
recipe	947 non-null	int64
calories	895 non-null	float64
carbohydrate	895 non-null	float64
sugar	895 non-null	float64
protein	895 non-null	float64
category	947 non-null	object
servings	947 non-null	object
high_traffic	574 non-null	object

Presence of null values

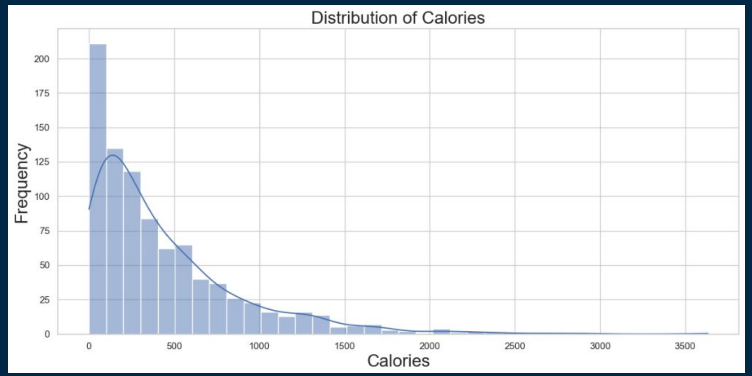
There was 11 values (there were supposed to be 10)

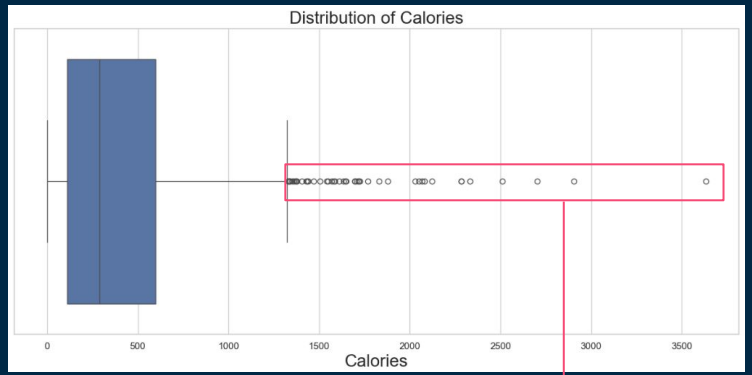
Wrong data type

Integer

Boolean

1. DATA EXPLORATION





All 4 numerical columns (Calories, Carbohydrate, Sugar and Protein) are right-skewed with a higher concentration in the values closer to 0.

The boxplots make evidence the existence of outliers.

We will use the median as a metric.
We assume there are outliers in the data.

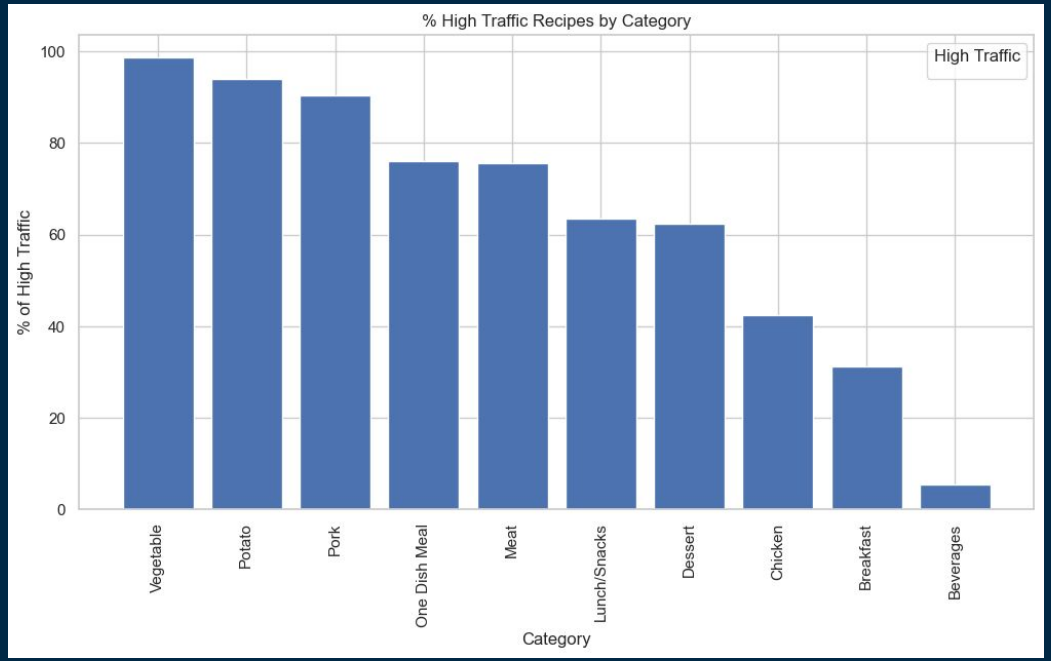
What can we do to further analyze this dataset?

Try to understand what role do the categories have towards the traffic.

Outliers

Apply Yeo-Johnson Transformation

1. DATA EXPLORATION



There are some categories that tend to showcase an increased traffic.

The top three categories in terms of high traffic are:

- Vegetable with a 99% of high traffic.
- Potato with a 94% of high traffic.
- Pork with a 91% of high traffic.

2. MODEL DEPLOYMENT

Main problem:
Determining if a recipe will showcase high traffic or not. ➡ This is a binary classification problem.

We have multiple classification models:
- Logistic Regression
- Decision Tree
- Random Forest
- Support Vector Machines

We choose:
- The Logistic Regression as the base model.
- The Support Vector Machines as the comparison model.

2. MODEL EVALUATION

We want to get an over 80% accuracy on determining if a post will show high traffic or not.

The best metric to compute this is Recall.

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Recall tells us the ratio of correctly predicted high traffic observations to the all observations in the high traffic class.

LOGISTIC REGRESSION MODEL:

```
Logistic Regression Train:
  Accuracy: 0.7472067039106145
  Recall: 0.8135198135198135
  Confusion Matrix:
[[186 101]
 [ 80 349]]
Logistic Regression Test:
  Accuracy: 0.770949720670391
  Recall: 0.839622641509434
  Confusion Matrix:
[[49 24]
 [17 89]]
```

- There is no overfitting.
- The slight elevation of test results implies the presence of limited data.
- **The recall metric was over 80%, so we achieve our objective.**

SUPPORT VECTOR MACHINES:

```
Support Vector Machines Train:
  Accuracy: 0.5991620111731844
  Recall: 1.0
  Confusion Matrix:
[[  0 287]
 [  0 429]]
Support Vector Machines Test:
  Accuracy: 0.5921787709497207
  Recall: 1.0
  Confusion Matrix:
[[  0  73]
 [  0 106]]
```

- Shows underfitting in the training set.
- The recall value is 1, indicating all elements are being labeled as high_traffic.

SVM ARE NOT WORKING.

3. HOW TO HELP THE BUSINESS?

The main goal of our business is to maximize the traffic in our website.

This means that we want to maximize the detection of high traffic recipes.

Having recall as a metric allows us to minimize the chance of missing out on high-traffic recipes.

Main Advantages:

- **Minimize missed opportunities:** The company aims to ensure that high-traffic recipes are not overlooked.
- **Operation Efficiency:** Increasing the number of predicted high-traffic recipes will increase the traffic in the website.

3. HOW TO HELP THE BUSINESS?

Main Recommendations:

- **Enhance data collection** to improve the model's accuracy and reliability, it is better to perform enhancements in data collection practices, particularly by including more granular details about user engagement with each recipe and feedback mechanisms.

- **The usage of a dashboard** to monitor real-time analytics that provides ongoing insights into which recipes are performing well and why. This tool would help the product team make data-driven decisions quickly and efficiently.

THANKS IN ADVANCE!