Traffic analysis from Recipes

Tasty Bytes

Introduction

The objective is to make a machine learning model that **predicts which recipes will lead to high traffic and correctly predict high traffic recipes 80% of the time**.

The results of this project will help guide the selection of recipes for display on the homepage, potentially leading to a significant **increase in website traffic and subscriptions**.

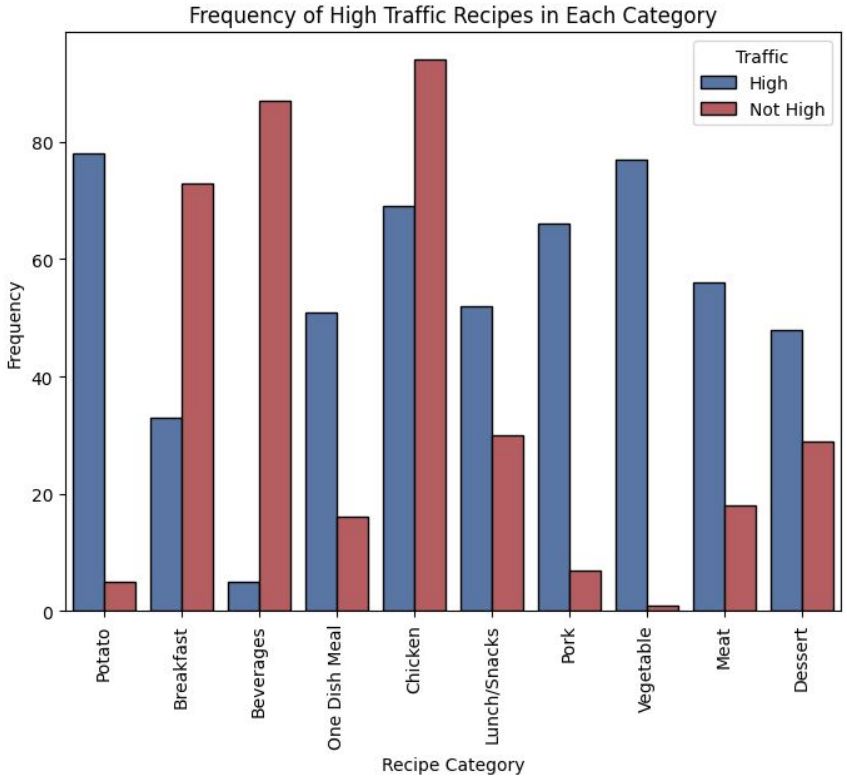
Dataset Overview

We **dropped** the rows with missing values in "calories", "carbohydrate", "sugar", and "protein" columns to **ensure complete and reliable data for analysis**. This only resulted in a loss of **5.5%** of the original data. Dropping the rows is a reasonable decision given the small percentage of missing data and the importance of having **complete data for the predictive model**.

Column Name	Details	Validation of Data Set and Cleaning Steps If Necessary
recipe	Numeric, unique identifier of recipe	895 non-null. No cleaning of data was necessary
calories	Numeric, number of calories	895 non-null. The values are numeric.
carbohydrate	Numeric, amount of carbohydrates in grams	895 non-null. The values are numeric.
sugar	Numeric, amount of sugar in grams	895 non-null. The values are numeric.
protein	Numeric, amount of protein in grams	895 non-null. The values are numeric.
category	Character, type of recipe. Recipes are listed in one of ten possible groupings (Lunch/Snacks', 'Beverages', 'Potato', 'Vegetable', 'Meat', 'Chicken, 'Pork', 'Dessert', 'Breakfast', 'One Dish Meal')	895 non-null. It was necesary to change the value "Chicken Breast" for Chicken" this was done to 94 instances.
servings	Numeric, number of servings for the recipe	895 non-null. I change 2 values "4 as a snack" for 4, and 1 value "6 as a snack" for 4 finally I change the type from 'object' to 'int64'
high_traffic	Character, if the traffic to the site was high when this recipe was shown, this is marked with “High”.	535 non-null and 360 NaN values, I replace the NaN values for "Not High"

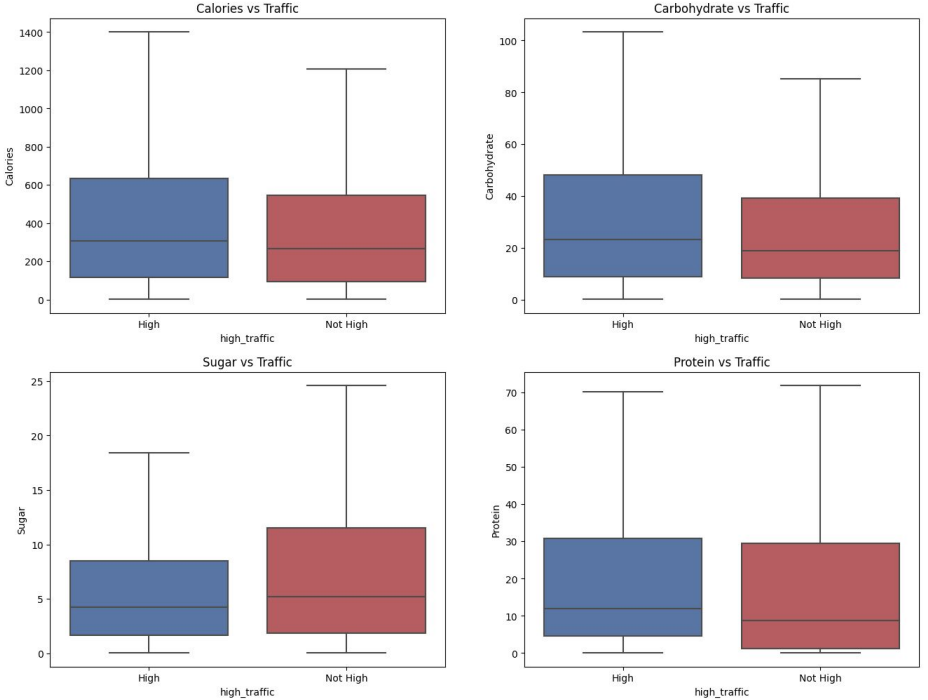
Recipe categories and popularity

We investigated the relationship **between recipe categories and popularity** by creating a plot of "category" vs "high_traffic" using a countplot. **Our observation indicated that certain categories, including "Potato", "Vegetable", and "Pork", had a higher frequency of high traffic recipes. Conversely, "Beverages" and "Breakfast" had a lower frequency of high traffic recipes.** This information can help the company identify areas for growth and develop new recipes to cater to the preferences of their website visitors.



Nutrient values and recipe popularity

We investigated the relationship between **nutrient values and recipe popularity** by creating box plots of "calories", "carbohydrate", "sugar", and "protein" vs "high_traffic". However, **we did not observe any significant differences in nutrient values between high traffic and non-high traffic recipes**. This suggests that **nutrient values may not be** a major factor in predicting recipe popularity.



Machine learning model

For this study, we selected two models for predicting recipe popularity: **logistic regression and random forest**.

Logistic regression is a statistical model that predicts binary outcomes by estimating the probability of an event occurring, in this case, whether a recipe will be popular or not.

On the other hand, **random forest**, is a machine learning model that creates multiple decision trees and combines their predictions to make a final prediction. It is particularly useful for handling **non-linear relationships** between variables and dealing with large datasets.

Model Evaluation

To evaluate the performance of the logistic regression and random forest models, we used **precision and recall** as our metrics. Precision measures the percentage of correctly predicted high traffic recipes out of all the recipes predicted as high traffic, while recall measures the percentage of actual high traffic recipes that were correctly identified by the model.

The l**ogistic regression model achieved a precision score of 0.76 and a recall score of 0.86**. This means that the model correctly predicted 76% of the recipes predicted as high traffic and correctly identified 86% of the actual high traffic recipes. **The random forest model, on the other hand, achieved a precision score of 0.74 and a recall score of 0.82**. These results suggest that both models are capable of predicting high traffic recipes with a relatively high degree of accuracy.

However, when we compare the two models, we see that **the logistic regression model outperformed the random forest model in terms of both precision and recall**. This indicates that **the logistic regression model is better suited for predicting recipe popularity in this case.** The higher performance of the logistic regression model may be due to its ability to **model linear relationships between the predictor variables and the response variable**, which may be more appropriate for the data we are working with.

Improving the model

To monitor the business goal of predicting high traffic recipes, we recommend regularly evaluating the model's performance on new data and **tracking the success rate of correctly predicting high traffic recipes 80% of the time**.

A hypothesis test was conducted to determine whether the model can correctly predict high traffic recipes 80% of the time. The **p-value of 0.92** indicated that there was not enough statistical evidence to suggest that the model's accuracy in predicting high traffic recipes is significantly better than 80%.

We proposed the p-value of the hypothesis test in addition to precision and recall, as the key metric for evaluating the improvement of the logistic regression model, the objective is to keep improving the model until the p-value is **less than 0.05**

Recommendations for Improvement

The study aimed to develop a machine learning model to **predict high traffic recipes** that can increase website traffic and subscriptions.

The **logistic regression model** achieved a precision score of 0.76 and recall score of 0.86, **outperforming the random forest model**.

To improve the model's performance, we propose **exploring to include additional variables** such as the cooking time and preparation difficulty.

Collecting data for a longer period to include seasonal trends and variations in recipe popularity could also help to improve the model.

The logistic regression model demonstrated promising performance, and the company should continue to **monitor and refine** the model to achieve the desired level of accuracy which we don't have currently (evaluated by hypothesis testing).