Data Science Report on predicting popular recipes

Lucas Bernal Espenberger
28.08.2023

Introduction

- Current situation: a recipe is arbitrarily chosen by the Product Team and displayed on the homepage. Some recipes can increase the website traffic by up to 40%. We call this scenario high traffic.

- Aim of the project: be able to predict which recipes will be popular 80% of the time while minimizing the chance of showing unpopular recipes

Exploratory Analysis

A pattern related to missing nutritional values and their potential influence on higher traffic was observed.

Missing Nutritional Values

	Nº of Recipes	% of Recipes
High Traffic	**39**	**75%**
Low Traffic	**12**	**25%**

Containing Nutritional Values

	Nº of Recipes	% of Recipes
High Traffic	**535**	**59.78%**
Low Traffic	**360**	**40.22%**

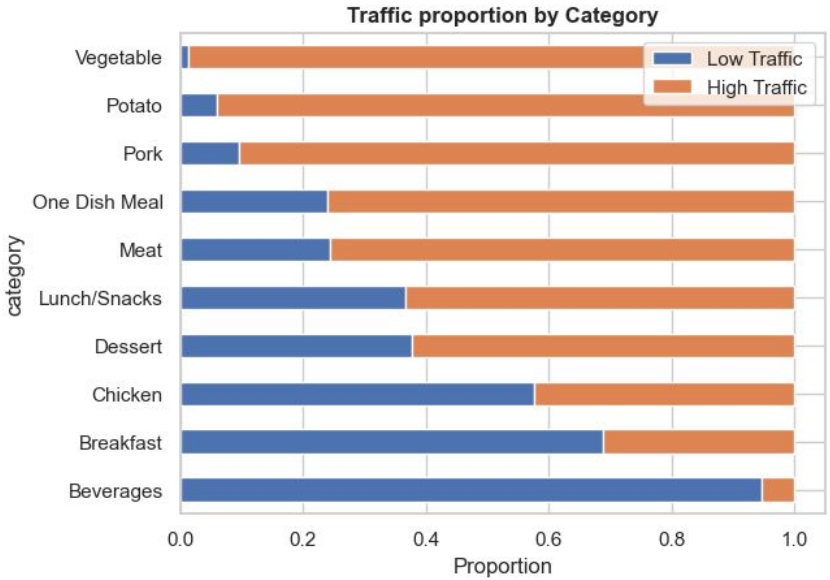
Statistical test of proportions between distinct samples:

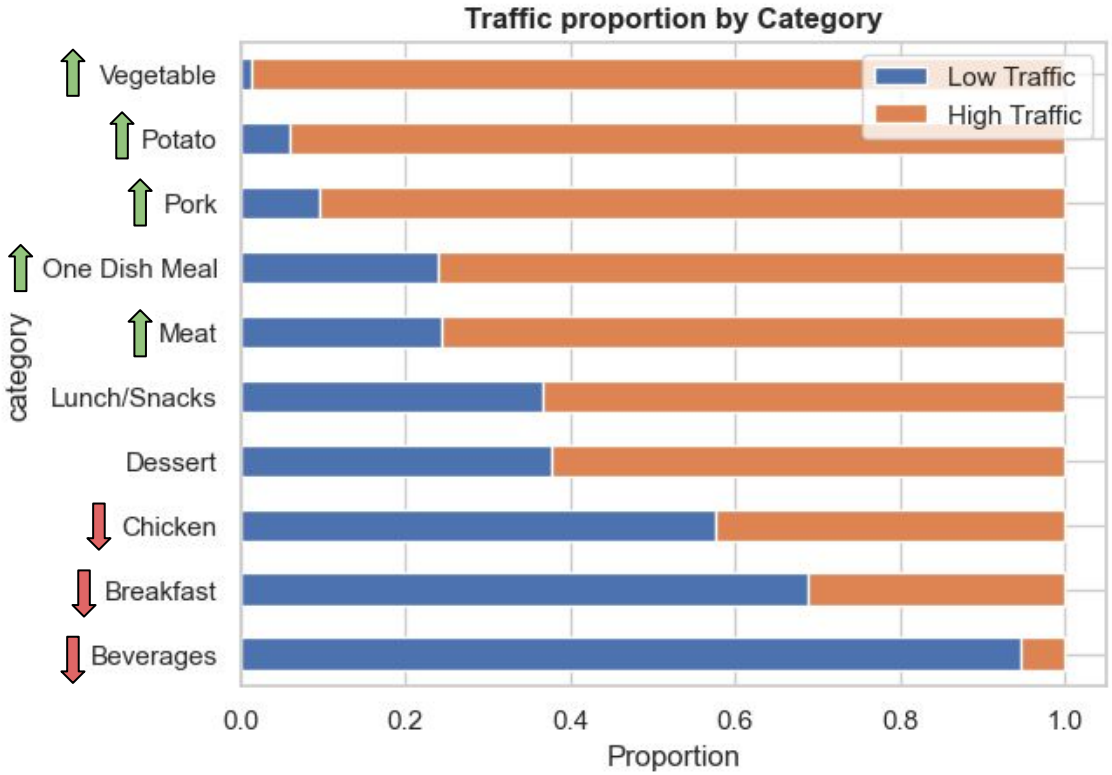
Recipes missing nutritional values have a significantly higher level of traffic than recipes containing nutritional values.

Exploratory Analysis

The proportions of high and low traffic recipes varied significantly among different categories.

A proportion test was conducted individually on each recipe category, comparing their traffic proportions to the overall sample proportion.





Proportion of high traffic recipes is significantly higher than the sample mean

Proportion of high traffic recipes is significantly lower than the sample mean

Predictive Model Development

"*They [Project team] want us to predict which recipes will be popular 80% of the time and minimize the chance of showing unpopular recipes.*"

Predictive model for binary classification

Best suited metrics to assess the performance of our predictive model:

- **Precision**: the proportion of true positive predictions (correctly predicted high traffic recipes) out of all the predicted positive instances (recipes predicted as high traffic).
- **False Positive Rate**: is the ratio of the number of false positives (recipes incorrectly predicted as high traffic) to the total number of actual negatives (actual low traffic recipes + false positives).

Baseline and comparison models

Baseline model: simple model that acts as a reference in a machine learning project.

- Majority Class Classifier: the model predicts the most frequent class in the training data for all instances in the test set.

Comparison model: more advanced predictive machine learning models

- Logistic Regression
- K-Nearest Neighbors

Results and Business metric

The metrics were chosen to align with the company's objectives of accurately predicting high-traffic recipes (high precision) while minimizing the display of low-traffic recipes (low FPR).

	Precision	False Positive Rate
Majority Class Classifier	60%	100%
Logistic Regression	78%	35%
K-Nearest Neighbors	81%	26%

The K-Nearest Neighbors model aligns with the company’s objective.

Summary and recommendations

- Recipes missing nutritional values have a significantly higher traffic than recipes that include their nutritional values.
- There is a significant correlation between recipe categories and website traffic dynamics.
- Predictive model that achieves a precision of almost 80% while yielding a low false positive rate was developed.