High traffic recipes prediction model

Introduction

Product Manager – Recipe Discovery asked for help to choose which recipes they should display at home page

A popular recipe drives high traffic to the website (upto 40% increase)

High traffic means more subscriptions

Their request:

Predict which recipe would lead to high traffic

An 80% correct prediction rate

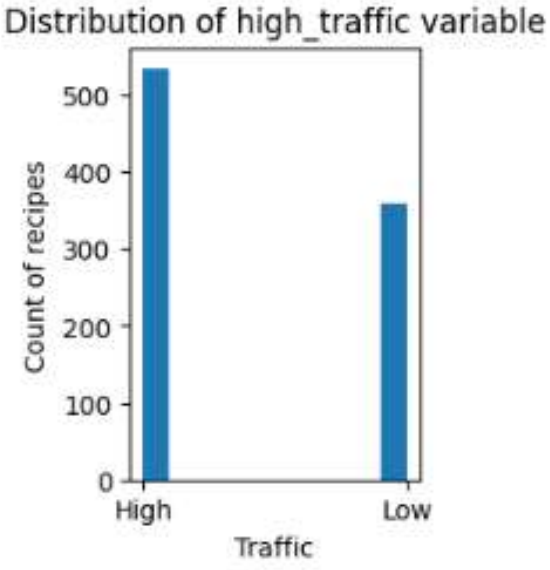
Data validation

Data set with 895 rows and 8 columns.

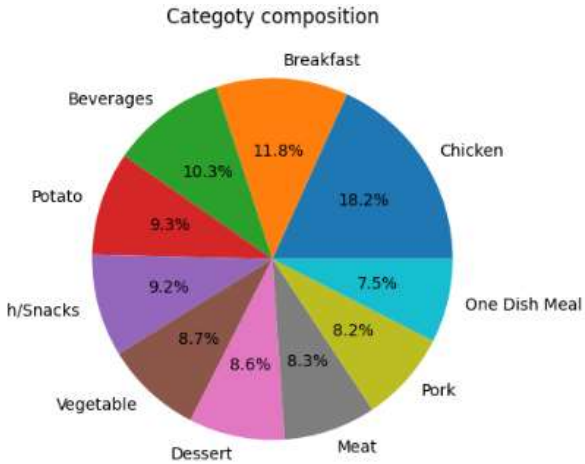
Recipe:	• No cleaning was needed.
Calories, Carbohydrate, Sugar, Protein:	• 52 missing values were found. They were removed
Category	• 11 categories were found of 10 possible groupings as indicated; an additional category "Chicken Breast" was found. This category was replaced with "Chicken" category
Servings:	• Some values were alphanumeric, they were cleaned, and data type was transformed to numeric as indicated.
High_traffic:	• 373 missing values were found, those values were replaced with a category named "Low"

Exploratory Data Analysis

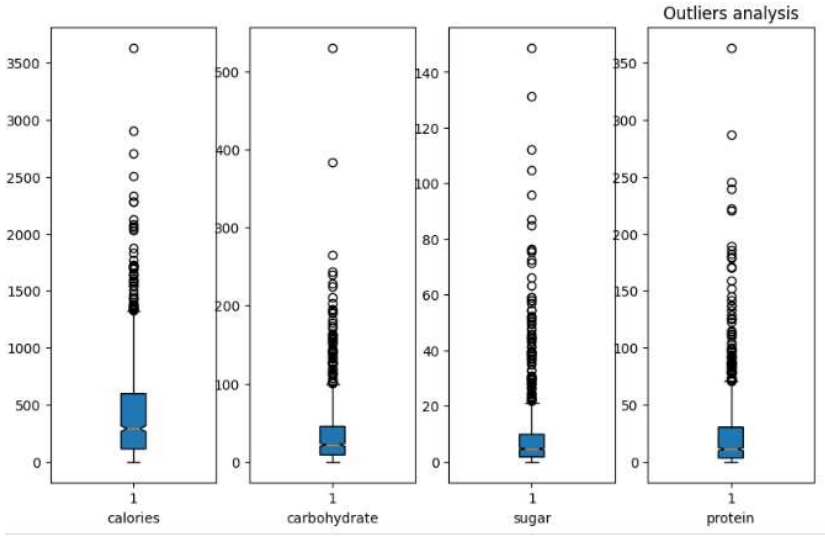
Single variable analysis



Class imbalance: <u>Acceptable</u>

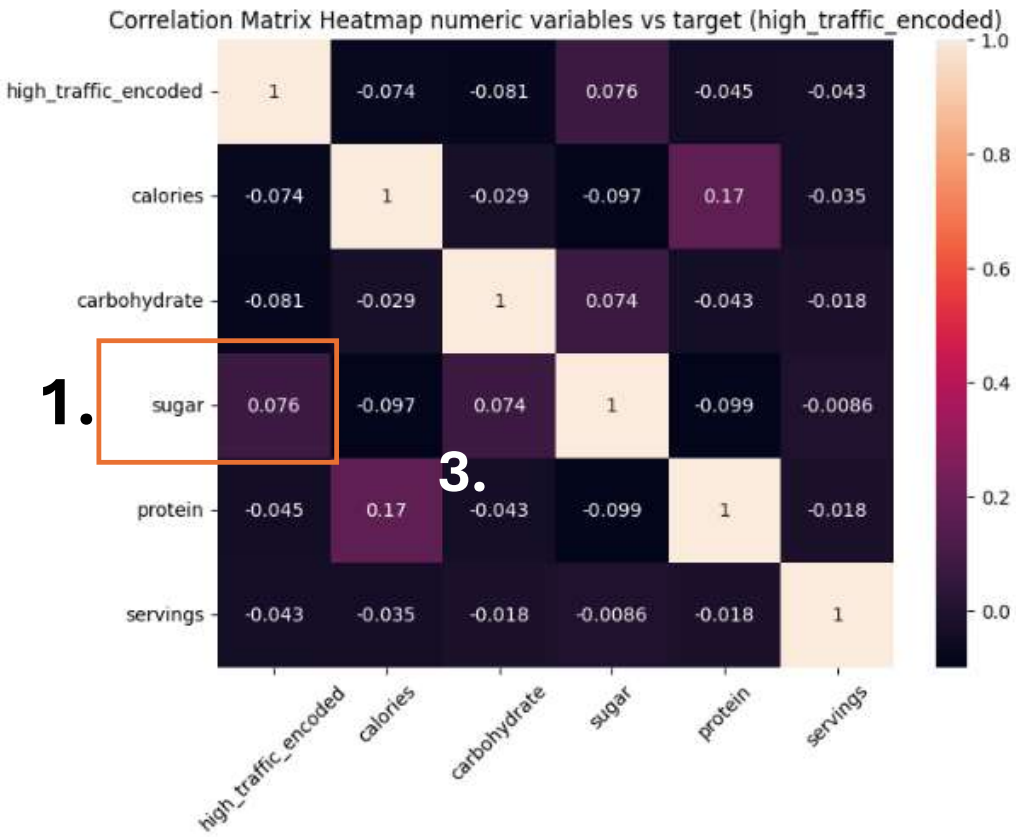


Most of the recipes are <u>Chicken</u>



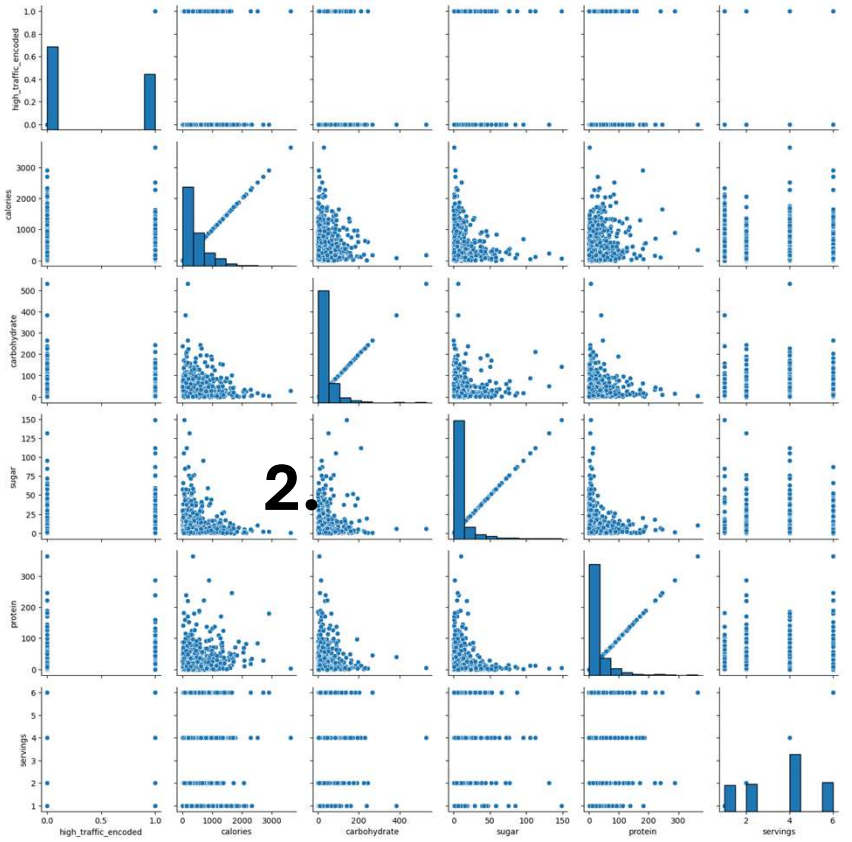
Outliers present in <u>calories, carbs, sugar and protein</u>

Cross-variables analysis



1.

3.



2.

1. Only Sugar variable has a positive relationship with high-traffic
2. No linear relationship between numeric variables and target
3. No highly correlated features (Multicollinearity)

Models & evaluation method

Identified problem: Classification

Classification Models:

Decision Tree

Random Forest

Evaluation Metric

Precision:

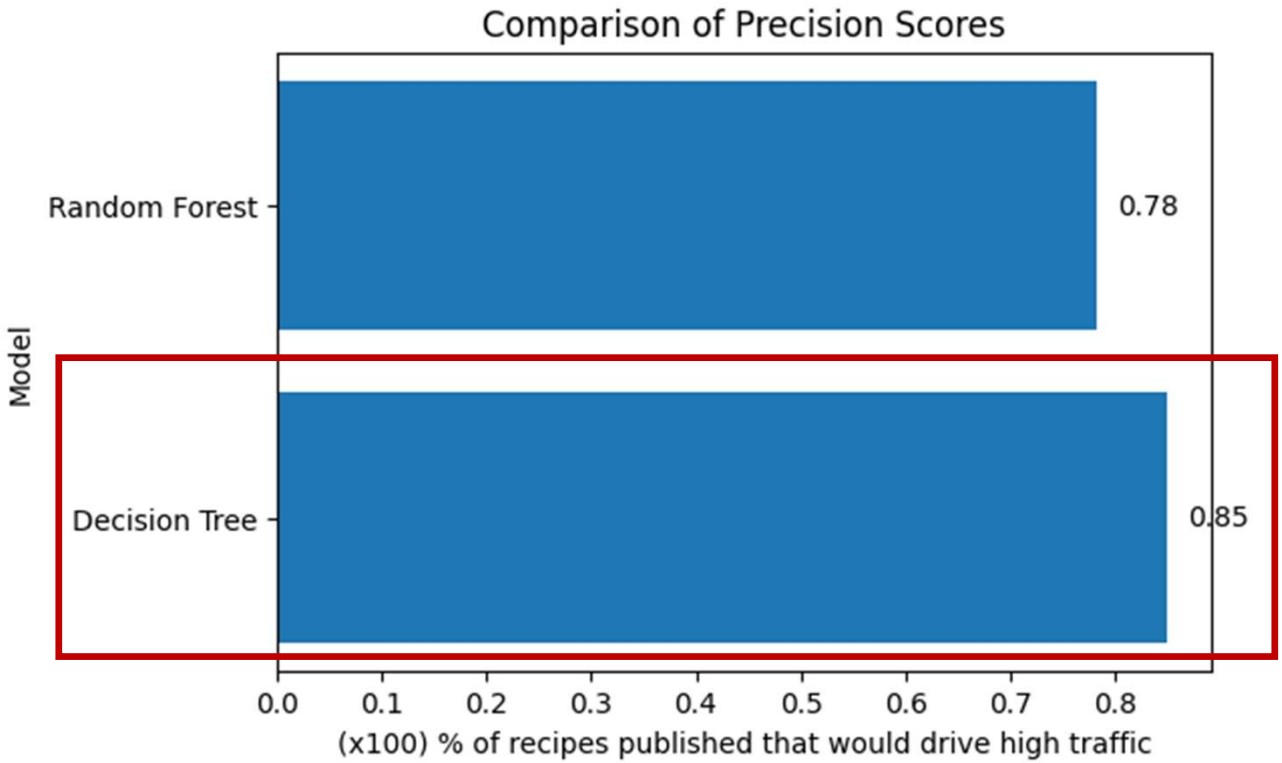
Correctly predicted as High traffic

All predictions as High Traffic

(# Correctly predicted as High traffic + # Wrongly predicted as High traffic)

Results

KPI: 80% of the publication of the recipe would drive an increase in the website traffic



Recommendations

- Deploy the Decision Tree Model into production with a user-friendly web service front-end
- Test the efficiency of the model in production in a two to three-month period
- Collect data through the usage of the model for retraining
- Measure the effectiveness of real results with the KPI defined.(80% of recipes drives high traffic)
- Experiment with including more features and compare the precision of the models