Tasty Bytes

Serving Up Success at Tasty Bytes

A data-driven approach to predict High-Traffic Recipes

Adrien Caudron

Feb 25th,2023

Executive Summary

- Context/Intro
- Data Overview
- Exploratory Data Analysis
- Feature Engineering
- Modeling Approach
- Model Performance
- Conclusions

Introduction

- Business Question:

Can we predict high-traffic recipes with at least 80% precision?

- Tasty Bytes
  - Founded in 2020 during the Covid-19 pandemic
  - Started as a recipe search engine to help people with cooking during the lock-down.
  - Now offers full meal plans for a monthly subscription, providing healthy and budget-friendly options for families.

- Project context and challenge
  - Tasty Bytes Product team has discovered that featuring a popular recipe on the homepage drives high traffic to the website.
  - We have historical data on website traffic related with recipe specificities.
  - Our challenge is to build a predictive model that can accurately identify recipes that will generate high traffic at least 80% of the time.

Data Overview (1/2)

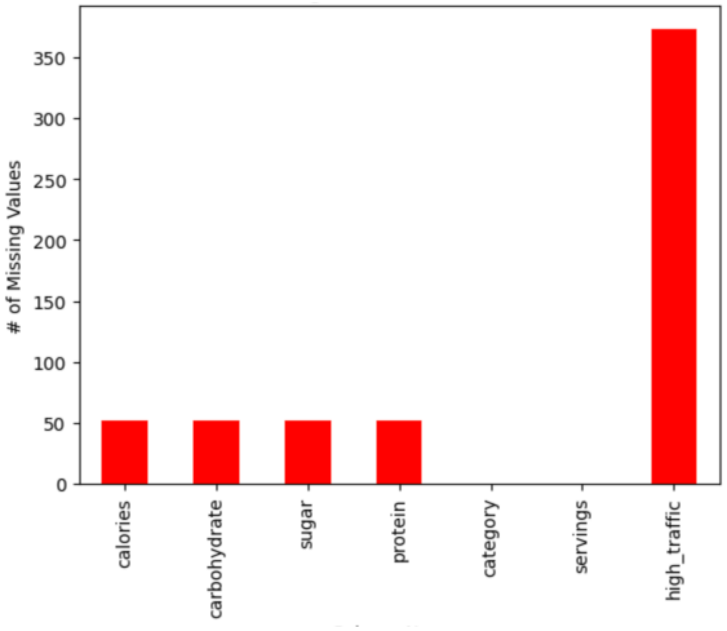
Quality Control / Data Transformation

- dataset columns : web traffic, calories, macro-nutrient content, number of servings, and category.

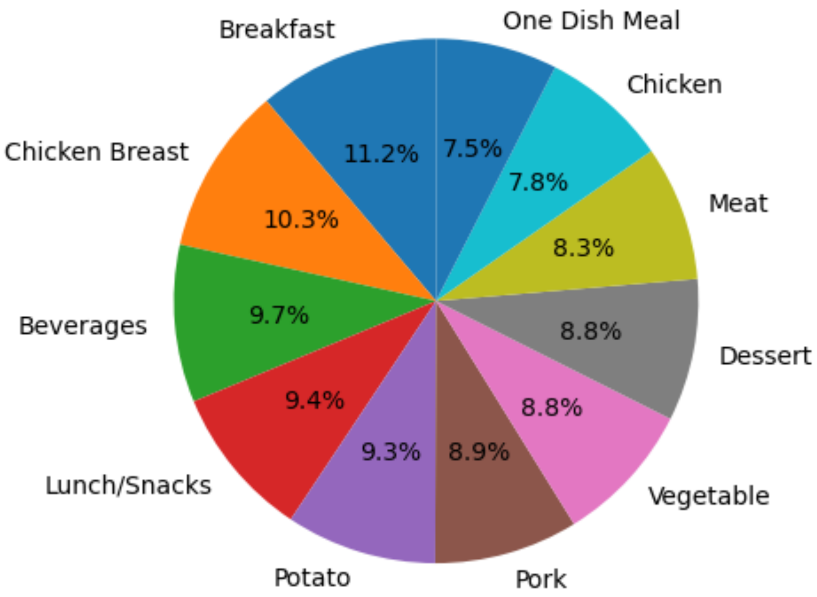
- The missing data in `calories` and macro-nutrient columns (5.5%) will be imputed using median values .

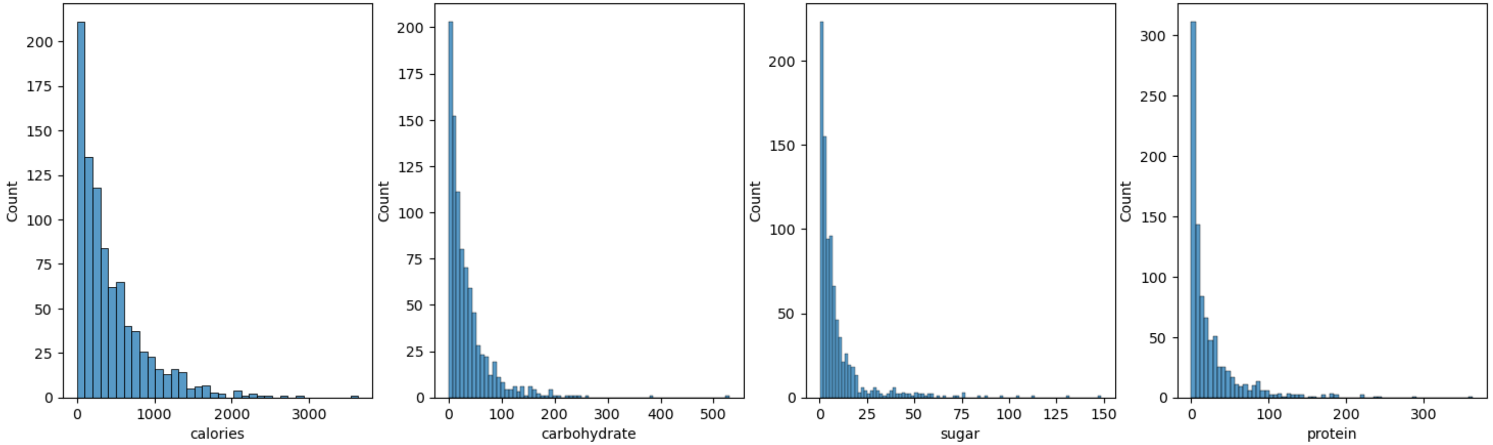
- The `category` column will be simplified by merging "`Chicken Breast`" and "`Chicken`” together.

Missing Values in our dataset



Percentage of Recipes per Category





Statistical distribution of calories and macro-nutrients columns

Data Overview (2/2)

Column Name	Variable Type	Data Type	Non-null count (% non-null)
high_traffic	Target (response)	Categorical	574 (60.7%)
calories	Feature (predictor)	Continuous (float)	895 (94.5%)
carbohydrates	Feature (predictor)	Continuous (float)	895 (94.5%)
sugar	Feature (predictor)	Continuous (float)	895 (94.5%)
protein	Feature (predictor)	Continuous (float)	895 (94.5%)
category	Feature (predictor)	Categorical	947 (100%)
servings	Feature (predictor)	Categorical	947 (100%)

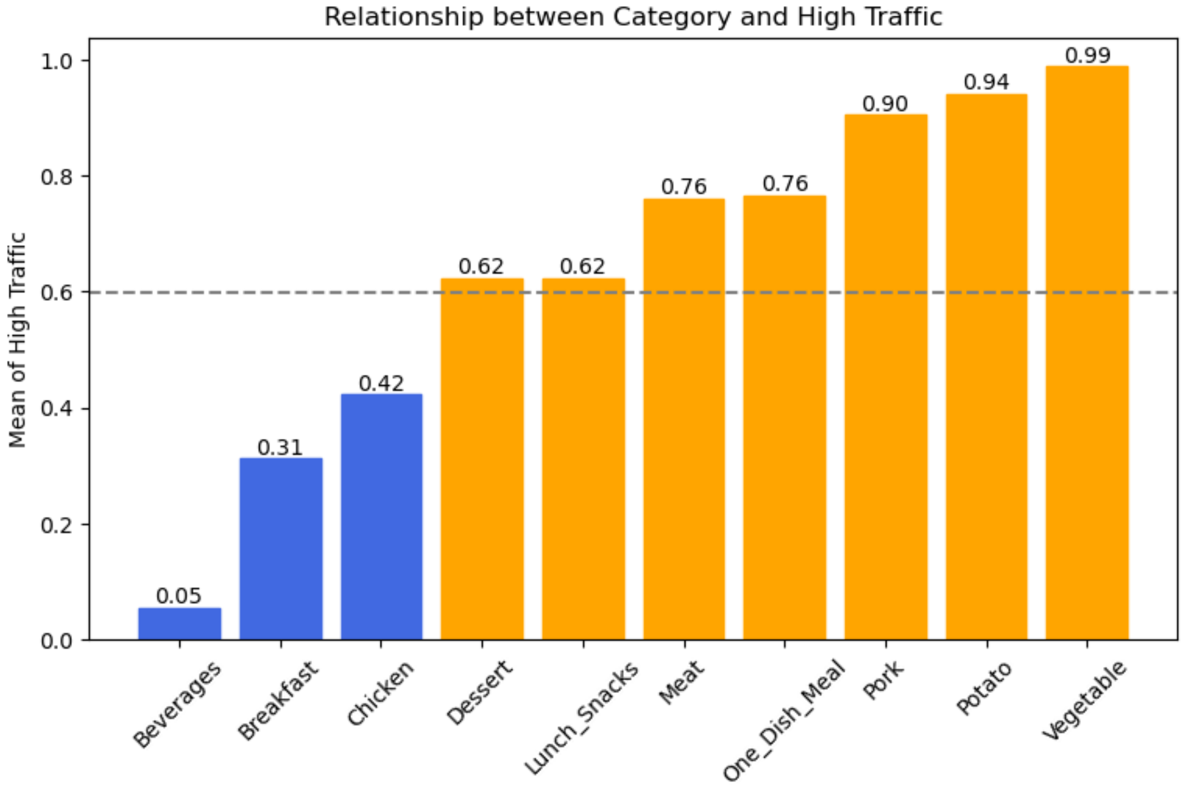
Statistical Significance and Predictability

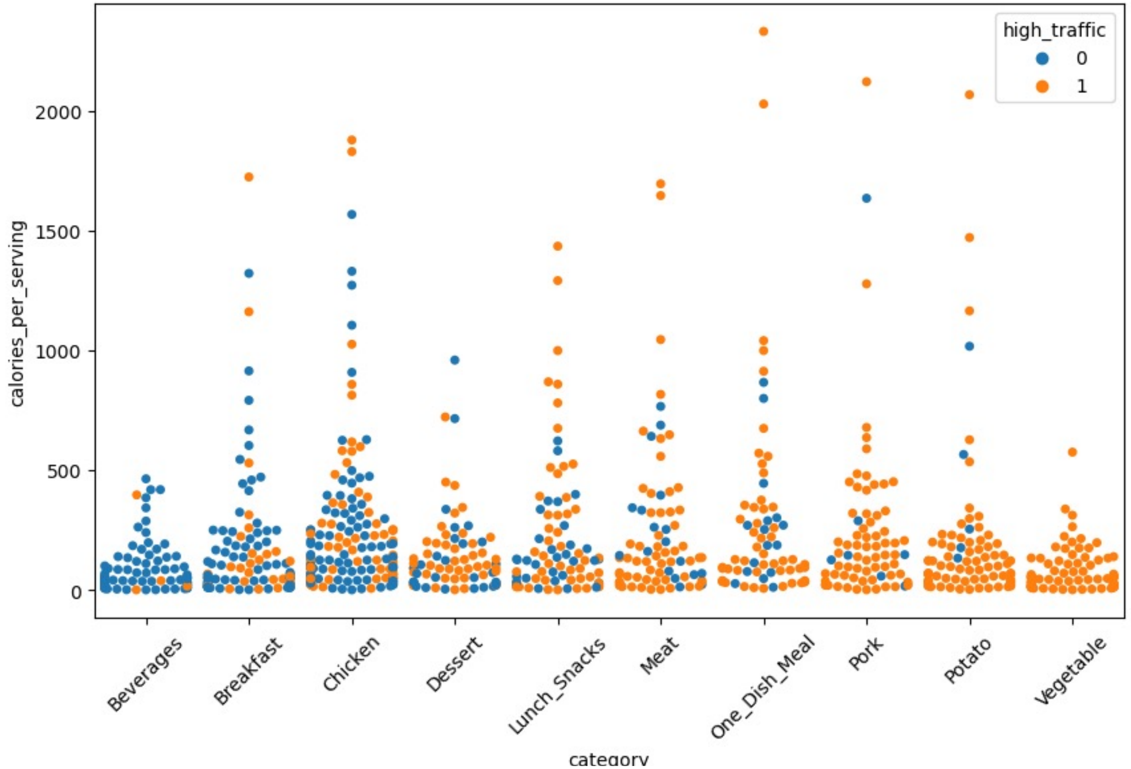
- `calories`, `carbohydrate`, and `sugar` columns are all statistically significant (p-values <0.5) when associated with the column `high_traffic`.
- Therefore, the null hypothesis that an association between those variables and `high_traffic` is likely to have occurred by chance alone can be rejected.
- **A feasibility test using a simple logistic regression** with only `calories`, `carbohydrate`, and `sugar` as predictors and `high_traffic` as target, **returned an accuracy of 59%, validating the predictive power of our dataset.**

Assumptions :

- **The `high_traffic` column will be the target variable**
- The missing data in the `high_traffic` column will be associated with 'Other' traffic, and the column will be transformed into Boolean data.
- Based on the above assumptions and observations, the historical dataset contains 574 cases of 'High' traffic and 400 cases of 'Other' traffic, making it suitable for performing binary classification using machine learning.

Exploratory Data Analysis (EDA)





- Rich meat and vegetable dishes (including potato) are more prevalent among Tasty Bytes users, possibly due to their filling and tasty nature and association with special occasions such as holidays or thanksgiving. This is coherent with Tasty Recipe's vocation of providing healthy and budget-friendly options for families
- Beverages, breakfast, and chicken-based meals are generating less traffic. They may be less popular due to their less unique taste and convenience factor. Chicken is considered lean meat and can be perceived as less filling. Beverages and breakfast could be simpler to realize, making the Tasty Bytes users less likely to look for those recipes.

→ Tasty Bytes indicates that users are likely seeking healthy and delicious food options.

Feature Engineering

Tasty Bytes

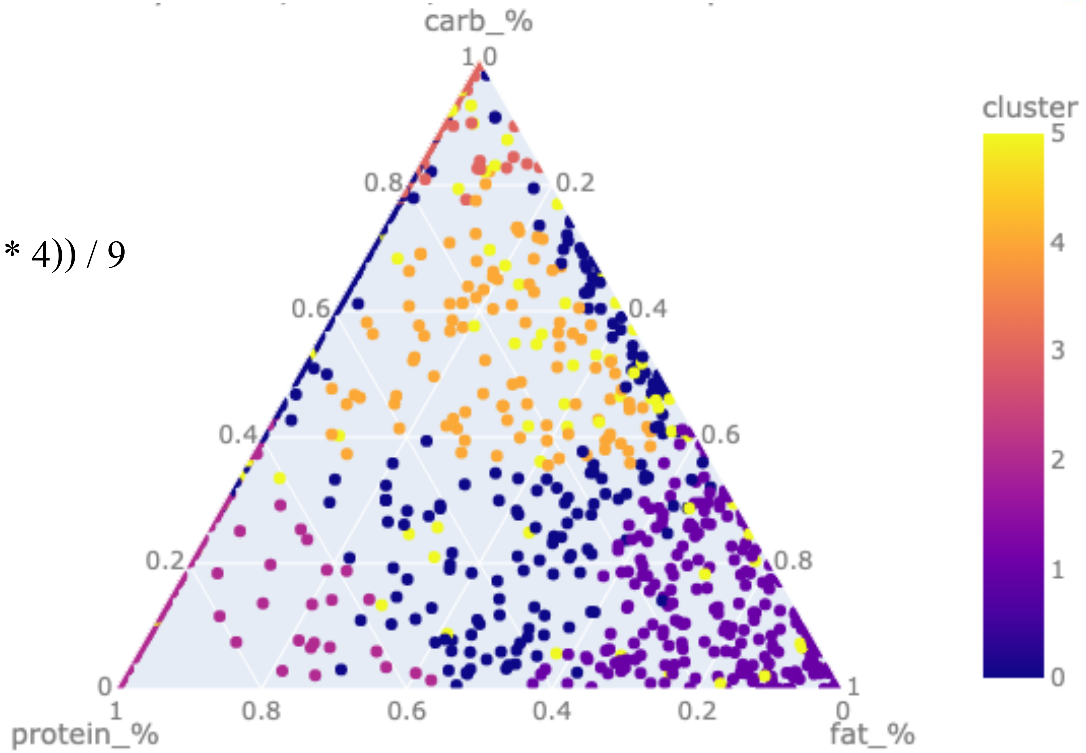
- We used a modified and simplified version of the Atwater system to estimate the grams of Fat from other macro-nutrient and calories columns :

Fat (g) = (Total Calories - (Protein (g) * 4 + Carbohydrates (g) * 4 + Sugar (g) * 4)) / 9

- We used the transformed servings columns to calculate the calories and macro-nutrients per serving.

- The Mayo Clinic recommends the following ranges for balanced foods:
  - Protein > 10% & Protein < 35%
  - Fat > 20% & Fat < 35%
  - Carb >= 45% & Carb < 70%
  - Recommended Daily intake for an adult: 1,600 calories/d
  - No more than 30g of sugar per day

- We adapted those cut-offs values to have slightly broader range and create 6 clusters of recipes



🟦 Type 0: Somewhat unbalanced foods
🟪 Type 1: High fat
🟪 Type 2: High protein
🟥 Type 3: High carb
🟧 Type 4: Healthier choices
🟨 Type 5 : Unhealthy recipe (high sugar)

Modeling Approach

- **Binary classification problem**: Our dependent variable, `high_traffic`, is a categorical variable that can only take one of two possible values, making binary classification the most appropriate approach for this problem.

- **Key Performance Indicators:**
  - **Precision: proportion of true positive predictions among all positive predictions.**
  - *Recall:* proportion of true positive predictions among all positive instances.
  - *AUC* (Area Under the receiver operating Curve) : measures how well the two classes are separated.
  - *F1 score:* harmonic mean of precision and recall

  → A model with a **precision score above 80%** will predict high-traffic recipes accurately 80% of the time.

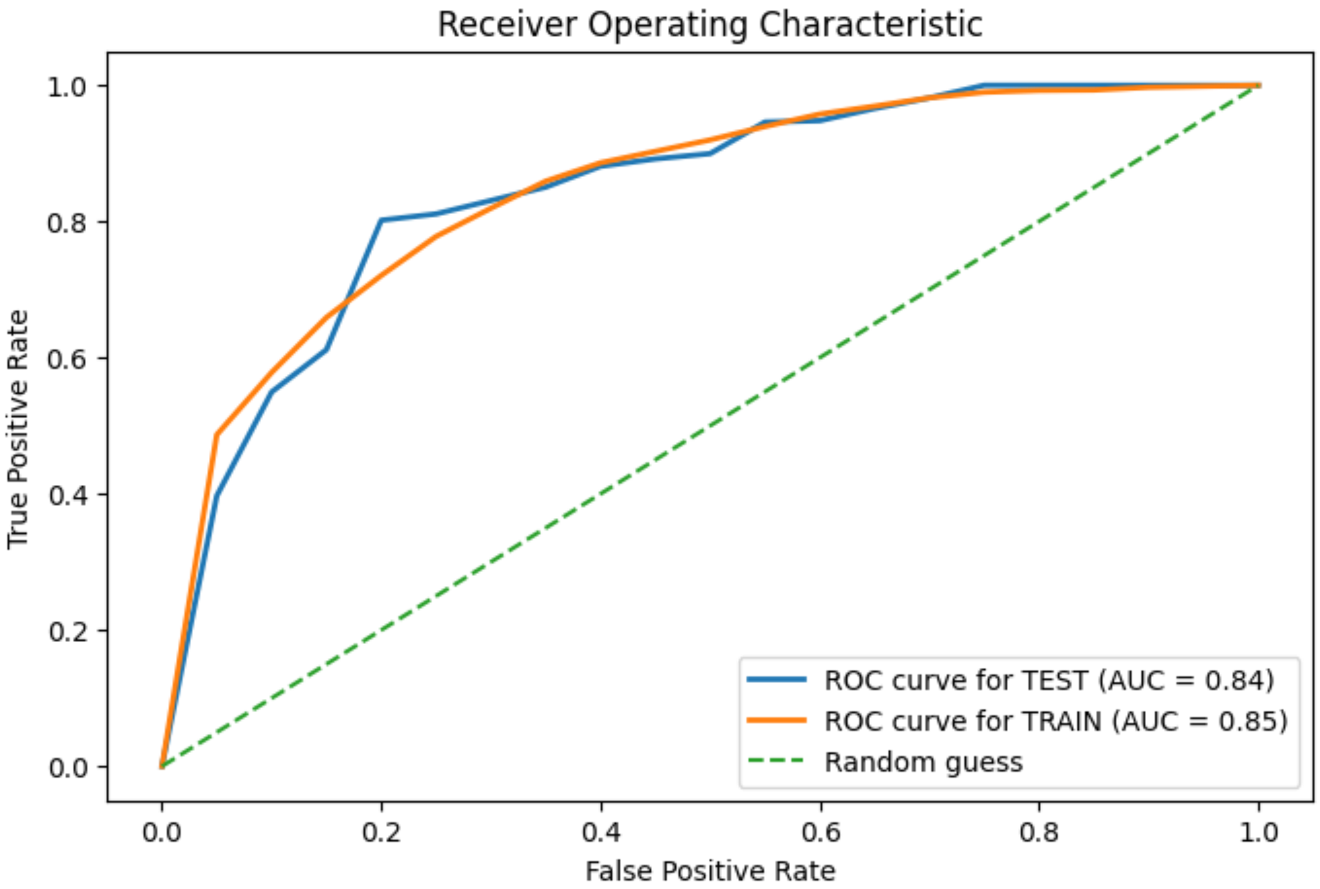
  F1, recall, and AUC scores above 80% are also desirable / Indicators of solid performance, limiting false negatives and adequate separation of our two categories.

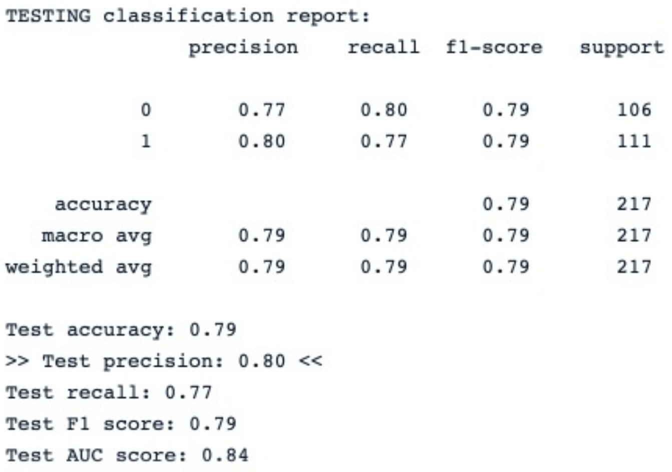
- **Preprocessing methods steps before Modeling:**
  - **Fixing Class Imbalance:** we used a SMOTE (Synthetic Minority Over-sampling Technique) approach to ensure we have the same number of values for the categories `High` and `Other` in the `high_traffic` target column.
  - **Converting numerical variables to bins:** `Fat_%` → `'fat_Low', 'fat_Mid’, 'fat_High'`
  - **Choosing the adequate features for predictive modeling.** We used the RFE (Recursive Feature Elimination) technique to keep only the 13 features with the highest contribution to our target variable: `high_traffic`
  -
  - Features selected: `'meal_Beverages', 'meal_Breakfast', 'meal_Chicken', 'meal_Dessert', 'meal_Pork', 'meal_Potato', 'meal_Vegetable', 'type_3', 'type_5', 'fat_Low', 'fat_Mid', 'protein_Low', 'sugar_High`

Modeling Performance

Baseline Model: Logistic Regression

Precision = 0.80





```
TESTING classification report:
              precision    recall  f1-score   support

           0       0.77      0.80      0.79       106
           1       0.80      0.77      0.79       111

    accuracy                           0.79       217
   macro avg       0.79      0.79      0.79       217
weighted avg       0.79      0.79      0.79       217

Test accuracy: 0.79
>> Test precision: 0.80 <<
Test recall: 0.77
Test F1 score: 0.79
Test AUC score: 0.84
```

Baseline : Logistic Regression

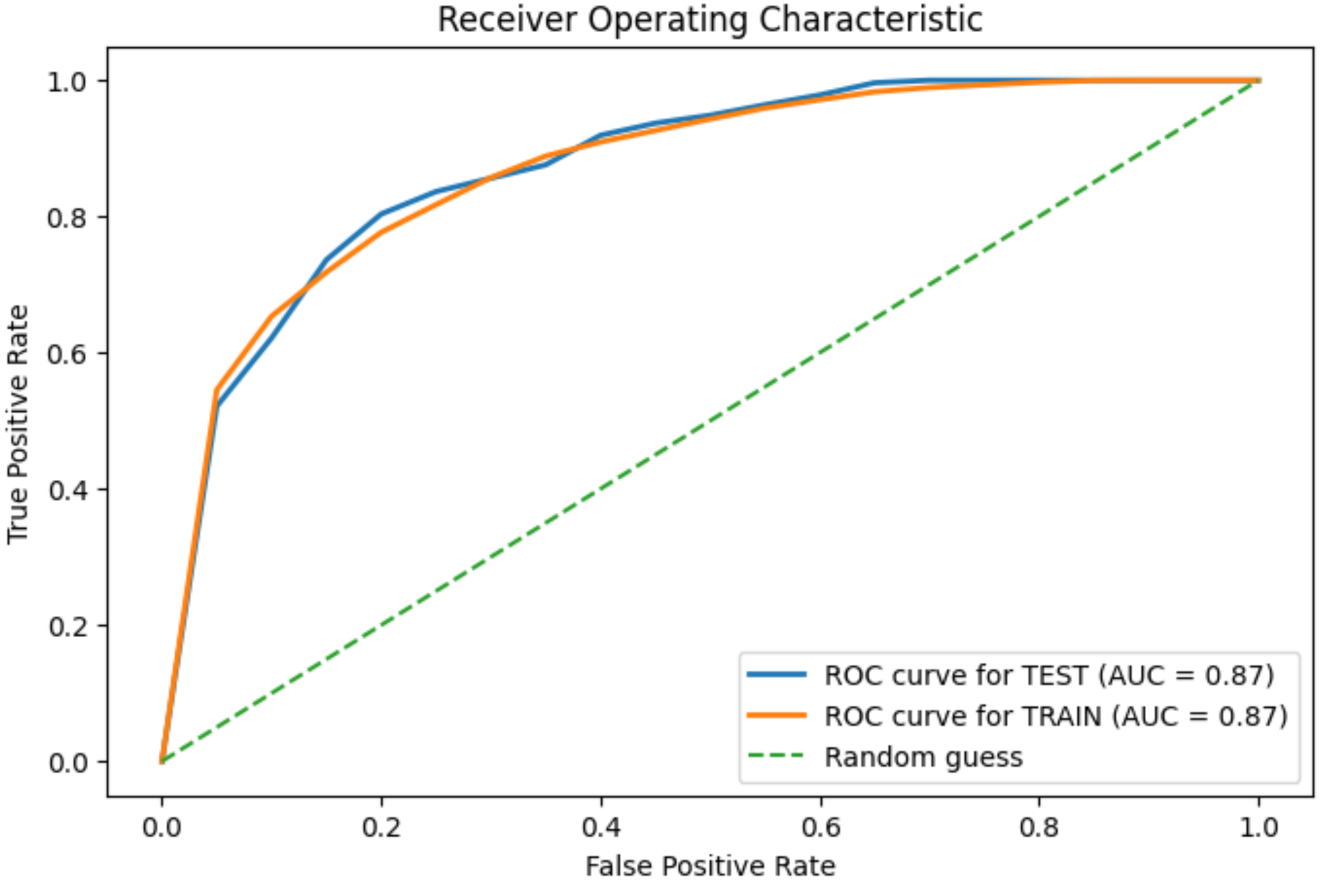
- Test precision: 0.80
- Test recall: 0.77
- Test F1 score: 0.79
- Test AUC score: 0.84

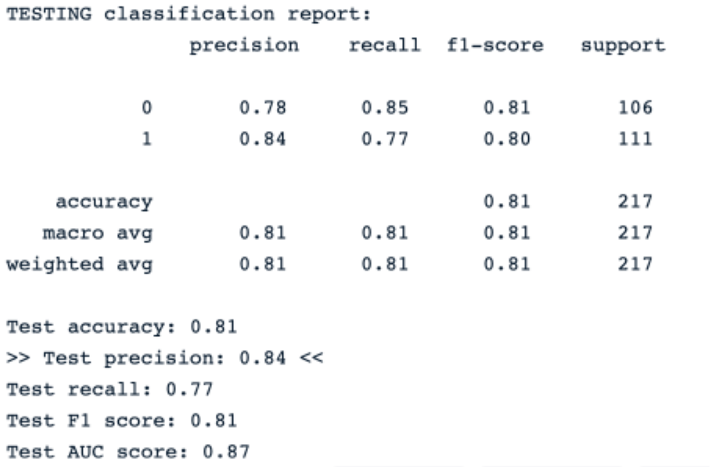
With a precision score above 80%, **our baseline model is already able to predict recipes generating high traffic at least 80% of the time.**

Modeling Performance

Best Model: CatBoost

Precision = 0.84





```
TESTING classification report:
              precision    recall  f1-score   support

           0       0.78      0.85      0.81       106
           1       0.84      0.77      0.80       111

    accuracy                           0.81       217
   macro avg       0.81      0.81      0.81       217
weighted avg       0.81      0.81      0.81       217

Test accuracy: 0.81
>> Test precision: 0.84 <<
Test recall: 0.77
Test F1 score: 0.81
Test AUC score: 0.87
```

Best Overall Model : CatBoost

- **Test precision: 0.84**
- Test recall: 0.77
- Test F1 score: 0.81
- Test AUC score: 0.87

With a precision score above 80%, we can claim that is able to predict recipes generating high traffic at least 80% of the time.

Note: We also tested 2 other high-performing models : SVM and XGboost (not displayed here).

Conclusions

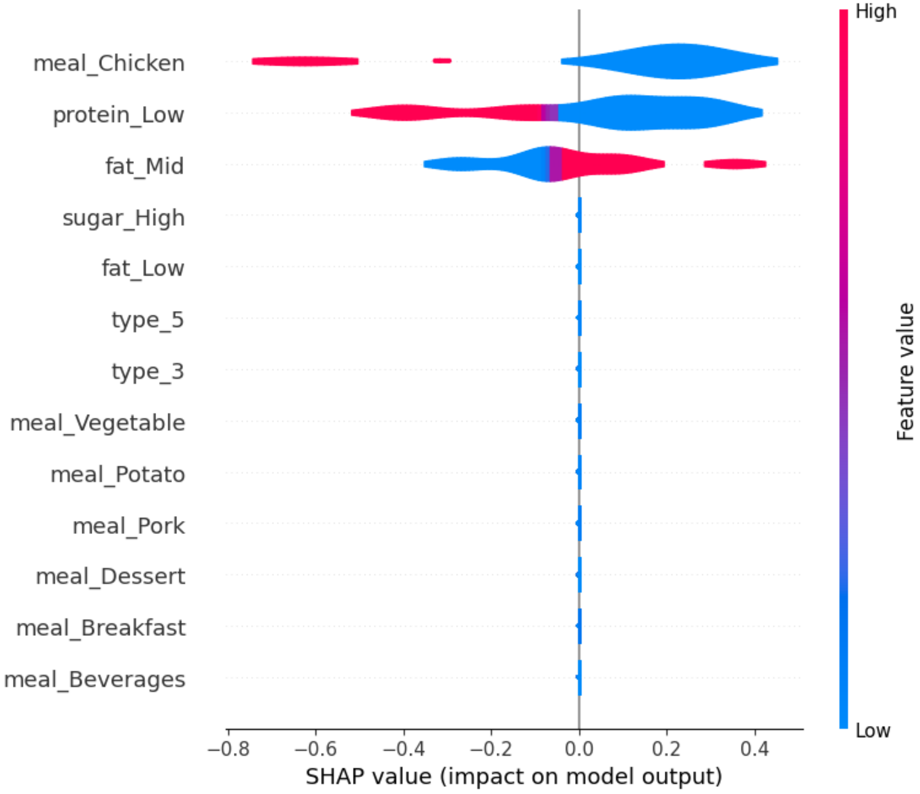
Tasty Bytes

- Our data enabled us to build predictive models that consistently predict recipes generating high traffic at least 80% of the time. Catboost, XGBoost, SVM, and Logistic Regression are our recommended algorithms for this task.
- Recommendations to further improve model performance:
  - Obtain more and higher-quality data
  - Improve feature engineering
  - Fine-tune machine learning models with more hyper-parameter tuning and cross-validation
  - Deploy models in production and implement MLOps to detect possible drift in new data.
- By implementing these recommendations, we can improve the accuracy and reliability of our predictive models and ensure that they continue to perform well over time.

Thanks!

Modeling Performance

Shapley Additive Explanations (SHAP) Analysis for CatBoost model



SHAP Results for CatBoost Model

Our analysis reveals that high traffic on the Tasty Bytes website is associated with recipes that are:

- NOT chicken-based
- NOT low in protein
- Moderately-High in Fat content.

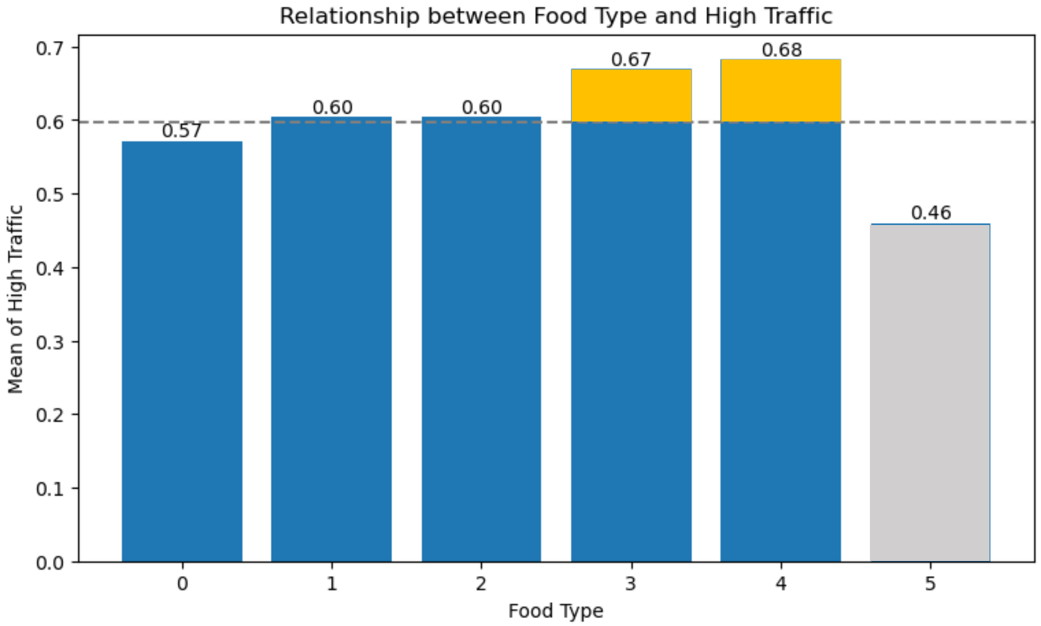
- Chicken dishes, low protein meals, and unbalanced fat content are associated with lower traffic and will likely negatively impact the website traffic and popularity.

- These findings corroborate our earlier observations:

Tasty Bytes visitors' primary interest lies in hearty and satisfying recipes suited for family gatherings.

Feature Engineering / EDA

Tasty Bytes



- Type 0: Somewhat unbalanced foods
- Type 1: High fat
- Type 2: High protein
- Type 3: High carb
- Type 4: Healthier choices
- Type 5: Unhealthy recipe (high sugar)

● **Type 3 (High-carb) recipes:** typically, pasta, bread, and rice are affordable and satisfying, making them a popular choice for people on a budget or with limited access to diverse ingredients.

● **(Type 4**) **Healthier choice recipes** could be popular due to being well-rounded and perceived as healthier.

○ **Type 5 (High sugar) recipes**, like desserts and fruity beverages, can be unpopular due to being perceived as unhealthy.

→ This resonates well with the reason d'être of Tasty Recipes, created during the pandemic as a search engine for recipes, helping people find ways to use up the limited supplies they had at home, and which now offer full meal plans for a monthly subscription, providing healthy and budget-friendly options for families.

Tasty Bytes

At the end of the data exploration analysis, we end up with those different columns:

- 'servings’
- 'high_traffic’
- 'calories_per_serving’
- 'carbohydrate_per_serving’
- 'sugar_per_serving’
- 'protein_per_serving’
- 'fat_per_serving’
- 'weight_per_serving’
- 'carb_%’
- 'protein_%’

'fat_%’

'sugar_%’

- 'daily_intake_%’
- 'meal_Beverages’
- 'meal_Breakfast’
- 'meal_Chicken’
- 'meal_Dessert’
- 'meal_Lunch_Snacks’
- 'meal_Meat’
- 'meal_One_Dish_Meal’
- 'meal_Pork’
- meal_Potato’
- 'meal_Vegetable’
- 'type_0’
- 'type_1’
- 'type_2’
- 'type_3’
- 'type_4’
- 'type_5’

- 'fat_Low'
- 'fat_Mid'
- 'fat_High’
- 'protein_Low’
- 'protein_Mid’
- 'protein_High’
- 'carb_Low’
- 'carb_Mid’
- 'carb_High’
- 'sugar_Low’
- 'sugar_Mid’
- 'sugar_High'