Predicting High-Traffic Recipes for Increased User Engagement

- Hamza Ali

Business Goals

Tasty Bytes, founded in 2020 during the Covid pandemic, began as a recipe search engine to help users make the most of limited supplies. Now a full-fledged business, Tasty Bytes offers monthly subscriptions that provide personalized meal plans for a healthy, balanced diet, with premium options including ingredient delivery to your door.

The business goals are to accurately predict and feature recipes on the company website's homepage that will:

- Drive high traffic to the website
- Increase user engagement and subscription rates.

By leveraging data-driven insights, the aim is to consistently identify popular recipes, optimize homepage content, and enhance overall user experience.

Project Overview

Problem Statement:

To meet the business goals stated earlier, the problems to be addressed are:

- Predict which recipes will lead to high traffic
- Correctly predict high traffic recipes at least 80% of the time

Approach:

The following approach was undertaken to address the problems:

- *Data Analysis*: Understand key features impacting recipe popularity.
- *Model Development*: Build and compare predictive models.
- *Evaluation*: Use metrics like Recall and F1 Score to assess model performance.

Data Analysis

Dataset Overview:

The dataset contained the features (variables):

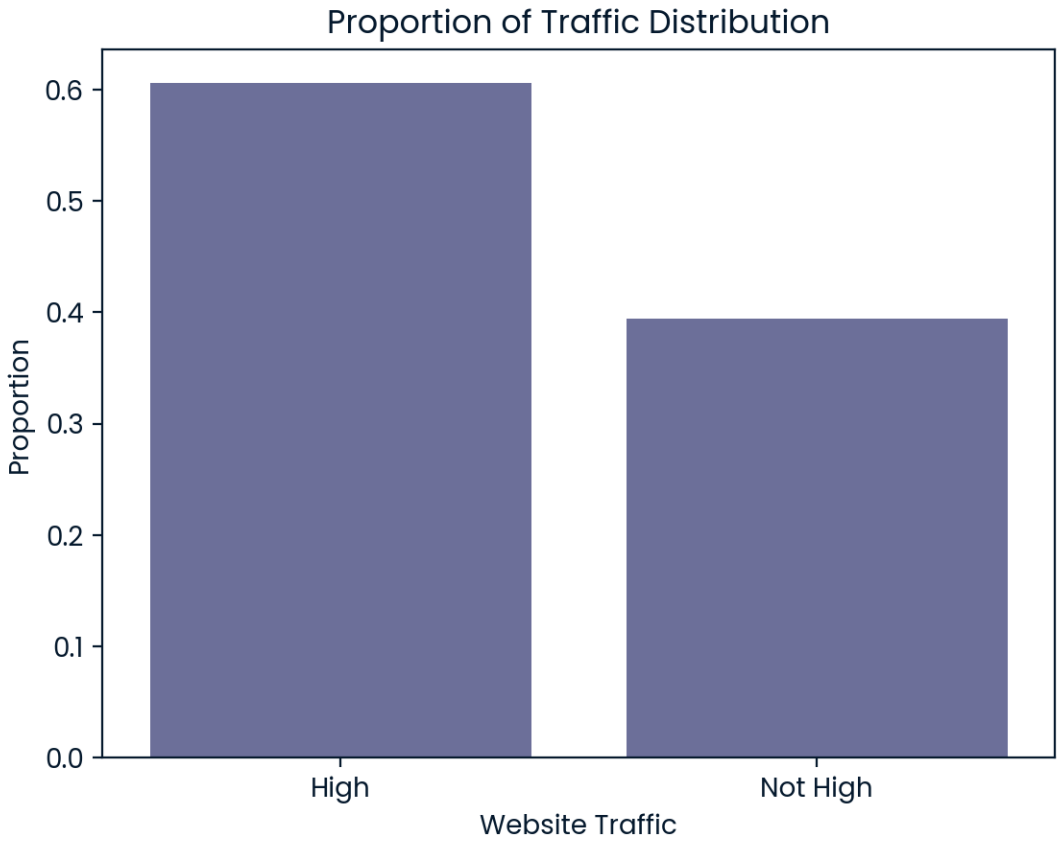
- **recipe**: A unique numeric identifier assigned to each recipe in the dataset.
- **calories**: The total number of calories per serving in the recipe, indicating its energy content.
- **carbohydrate**: The amount of carbohydrates (in grams) per serving in the recipe, reflecting its carbohydrate content.
- **sugar**: The amount of sugar (in grams) per serving in the recipe, indicating the sugar content.
- **protein**: The amount of protein (in grams) per serving in the recipe, showing the protein content.
- **category**: The type of recipe, categorized into one of ten possible groupings such as 'Lunch/Snacks', 'Desserts', 'Vegetable', etc.
- **servings**: The number of servings that the recipe yields, indicating the portion size.
- **high_traffic**: A categorical variable indicating whether the recipe led to high traffic on the website when it was featured, marked as "High" for high traffic and a missing value otherwise.

Data Cleaning and Validation:

Most of the features contained missing values, as well as certain values inconsistent with the original description of the feature, hence the dataset underwent a thorough cleaning procedure and was made ready for analysis.

Key Findings

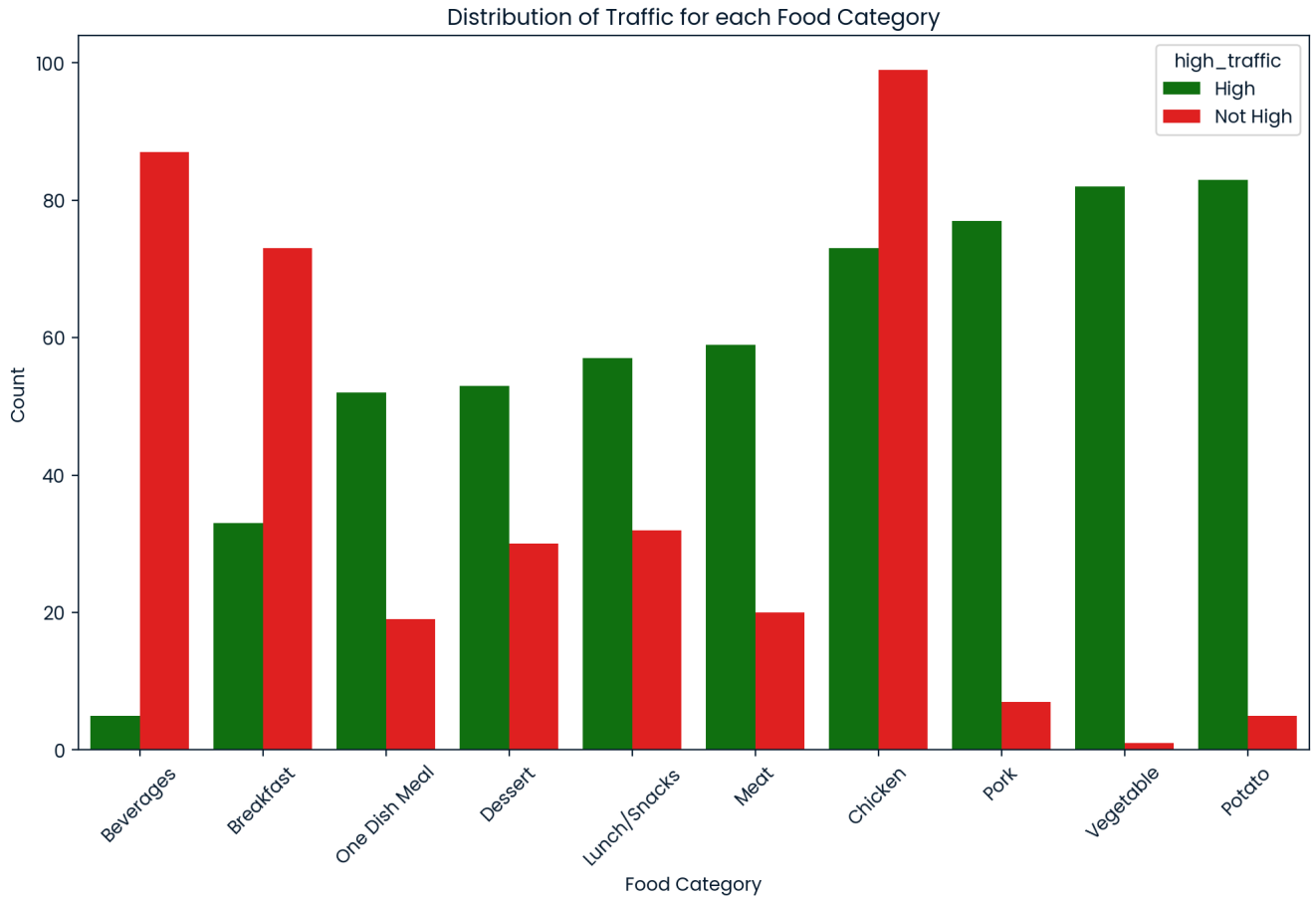
Website Traffic Proportion



The food recipes collectively accounted for almost 60% of the high traffic on the website, and about 40% of non-high traffic.

Key Findings

Website Traffic Proportion for each Food Category



The plot below shows how the website traffic is distributed for each food category. It shows that the food categories 'Potato', 'Vegetable' and 'Pork' are the most popular ones, while 'Beverages', 'Breakfast' and 'One Dish Meal' being the least popular.

Model Development and Evaluation

Data Preprocessing:

- Standardized numeric features.
- Encoded categorical features using one-hot encoding

Model Development:

- **Baseline Model**: Logistic Regression
- **Comparative Model**: Random Forest

Evaluation Metrics:

- **Recall**: Ensures most high-traffic recipes are identified
- **F1 Score**: Balances precision and recall for overall prediction quality.

Results:

The baseline and the comparative model both have similar performance in terms of recall, but the baseline or Logistic Regression model outperforms the comparative model or Random Forest model in terms of F1 Score. Another important thing to note is that the logistic regression model wasn't fine-tuned, but the Random Forest model was fine-tuned. Surprisingly, the baseline model, Logistic Regression model, is the model of choice.

Business Metrics

Key Metrics to Monitor:

- **Recall (0.82)**: Ensures 82% of high-traffic recipes are identified, maximizing opportunities for user engagement and revenue.
- **F1 Score (0.81)**: Balances recall with precision, minimizing the risk of featuring less popular recipes.

Impact on Decision-Making:

- **Recipe Selection**: High recall drives user engagement by ensuring popular recipes are featured.
- **Performance Monitoring**: Regularly track these metrics to adjust the model and maintain optimal website traffic.

Monitoring Strategy:

- **Regular Tracking**: Monthly evaluation of recall and F1 score.
- **Model Adjustment**: Fine-tune decision thresholds and retrain as needed.

Recommendations

To maximize site traffic and engagement, the business should:

- Fine-tune the logistic regression model to improve performance and metrics.
- Implement the Logistic Regression Model by deploying it on the website, for selecting homepage recipes.
- Retrain the Model with new data to adapt to trends.
- Monitor Recall and F1 Score regularly to maintain performance.
- Incorporate User Feedback for refining predictions.
- The analysis, indicates that the food categories 'Potato', 'Vegetable' and 'Pork' are the most popular ones and result in high traffic, so these should be used mostly as the homepage recipes.

These actions will help sustain high user engagement and increase subscription rates.

Thank You!