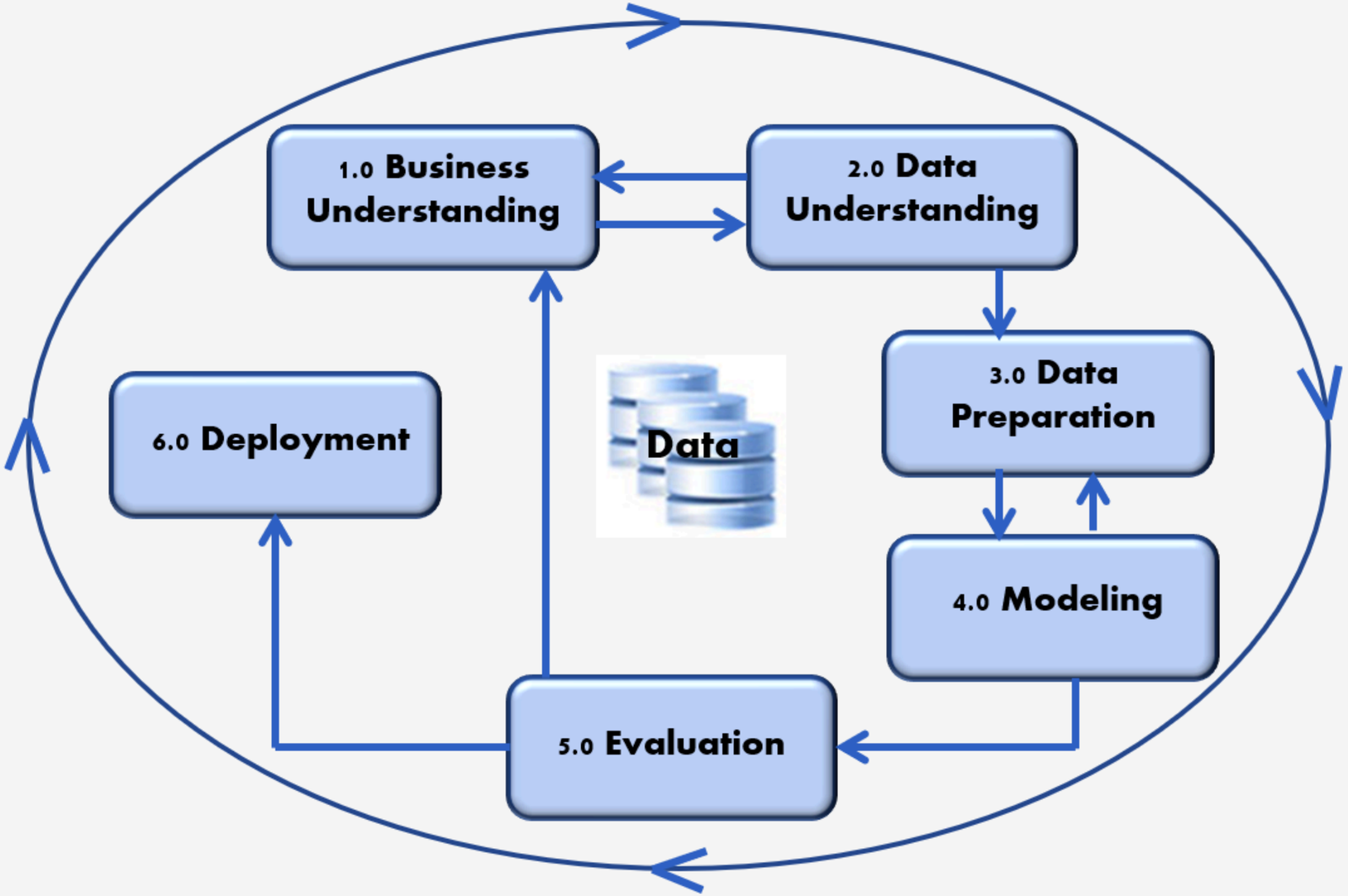
MAXIMIZING ENGAGEMENT: A DATA-DRIVEN APPROACH TO RECIPE SELECTION AT TASTY BYTES

JACOB MAXIMUS USARAGA

Problem Statement

Picking a **popular recipe** to be displayed on the home page generates an **uptick in traffic** in the rest of the website (as much as 40%). This leads to **more subscription.** We still have to find out how to predict **which recipes lead to high traffic**. We want to predict high traffic recipes **80%** of the time while **minimizing** the chance of **showing unpopular recipes.**

Data Science Methodology



Understanding, Preparing, and Cleaning data

Column Name	Details
recipe	Numeric, unique identifier of recipe
calories	Numeric, number of calories
carbohydrate	Numeric, amount of carbohydrates in grams
sugar	Numeric, amount of sugar in grams
protein	Numeric, amount of protein in grams
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').
servings	Numeric, number of servings for the recipe
high_traffic	Character, if the traffic to the site was high when this recipe was shown, this is marked with "High".

Handling Missing Information

We looked at our data and found that some information was missing in columns related to calories, carbohydrates, sugar, and protein. We then dropped the rows with missing values in the calories, carbohydrate, sugar, protein (5% of the data but it is necessary to drop it since there are missing values in the 4 columns) . Since there are no more missing values, we can proceed with the next steps.

Cleaning Data

Cleaned up the servings column by removing any unnecessary text and converting it into a number data type.

Converted the Chicken Breast Category to Chicken, since they are of the same category.

Understanding, Preparing, and Cleaning data

```
Data columns (total 8 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   recipe        895 non-null    int64
 1   calories      895 non-null    float64
 2   carbohydrate  895 non-null    float64
 3   sugar         895 non-null    float64
 4   protein       895 non-null    float64
 5   category      895 non-null    category
 6   servings      895 non-null    int32
 7   high_traffic  895 non-null    bool
dtypes: bool(1), category(1), float64(4), int32(1), int64(1)
```

Adjusting the Target Column

Updated the "high_traffic" column to use simple binary values (True and False) instead of text. The column indicates whether a recipe is expected to attract high traffic, which is what we want to predict. All other columns, excluding the recipe column, are used as features to help with this prediction.

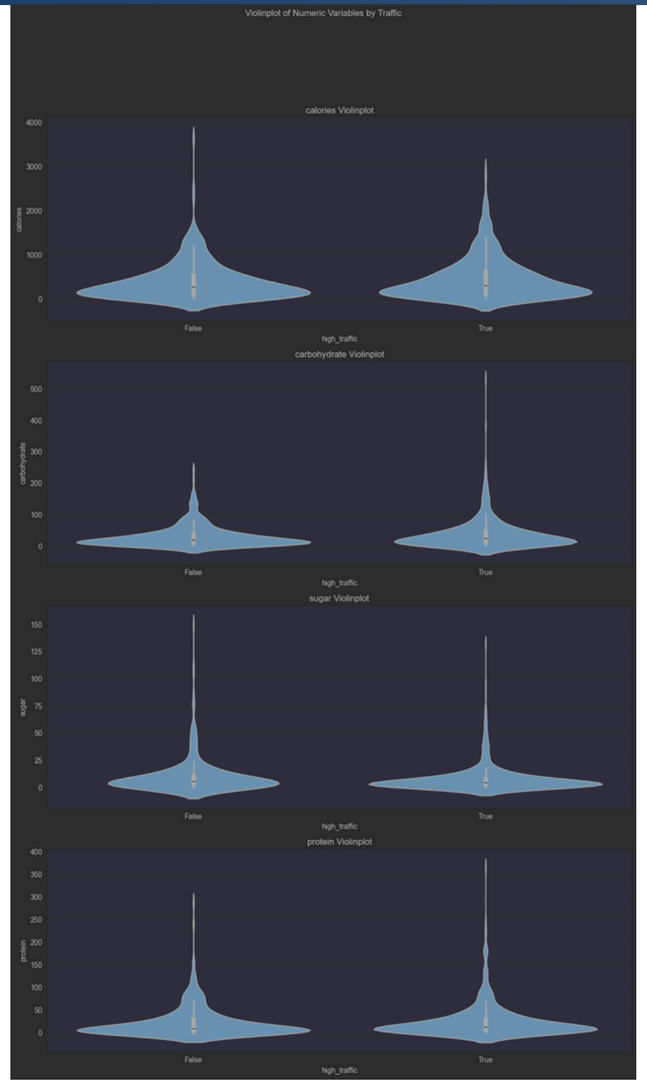
Data Retention

52 rows where removed from the 947 original rows of data. These rows have missing values in the protein, calories, carbohydrate and sugar column.

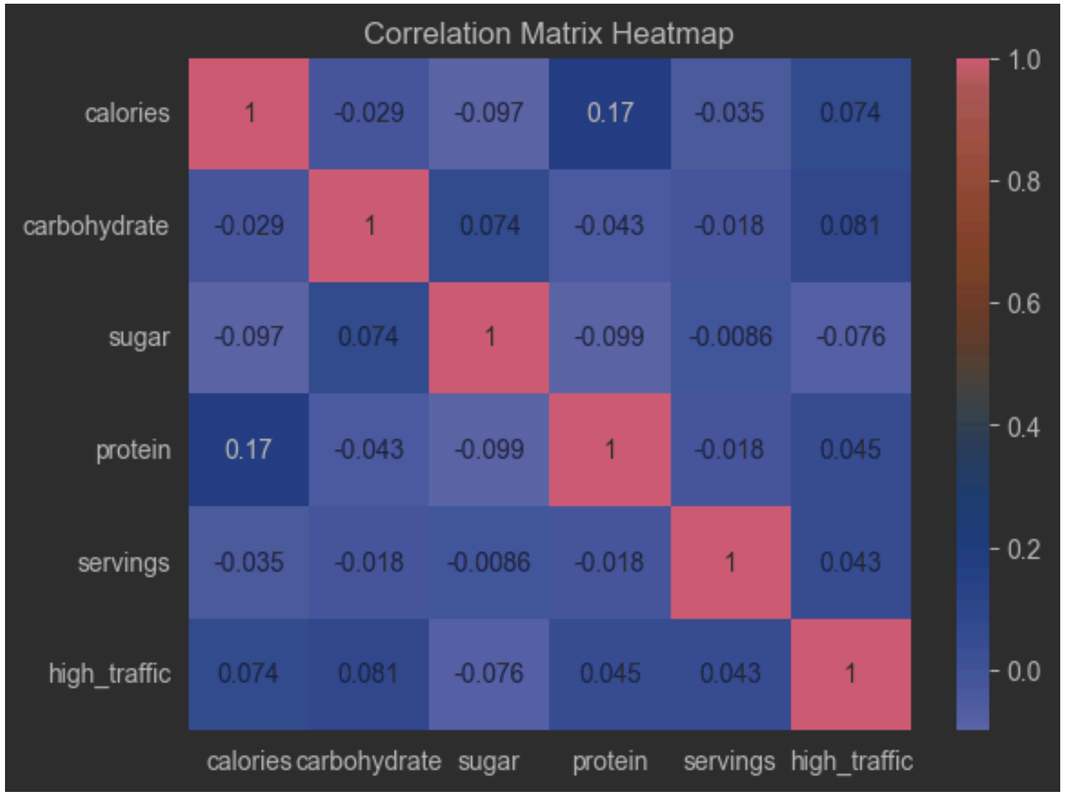
Analyzing and visualizing data

Distribution Similarity / Low Correlation Coefficient

The distributions of numerical variables (calories, carbohydrates, and protein) are similar between the two categories of the high_traffic target variable (true vs. false). This indicates that these variables might not significantly differentiate between high-traffic and non-high-traffic recipes, suggesting they may not be very effective for distinguishing between these categories.



Analyzing and visualizing data

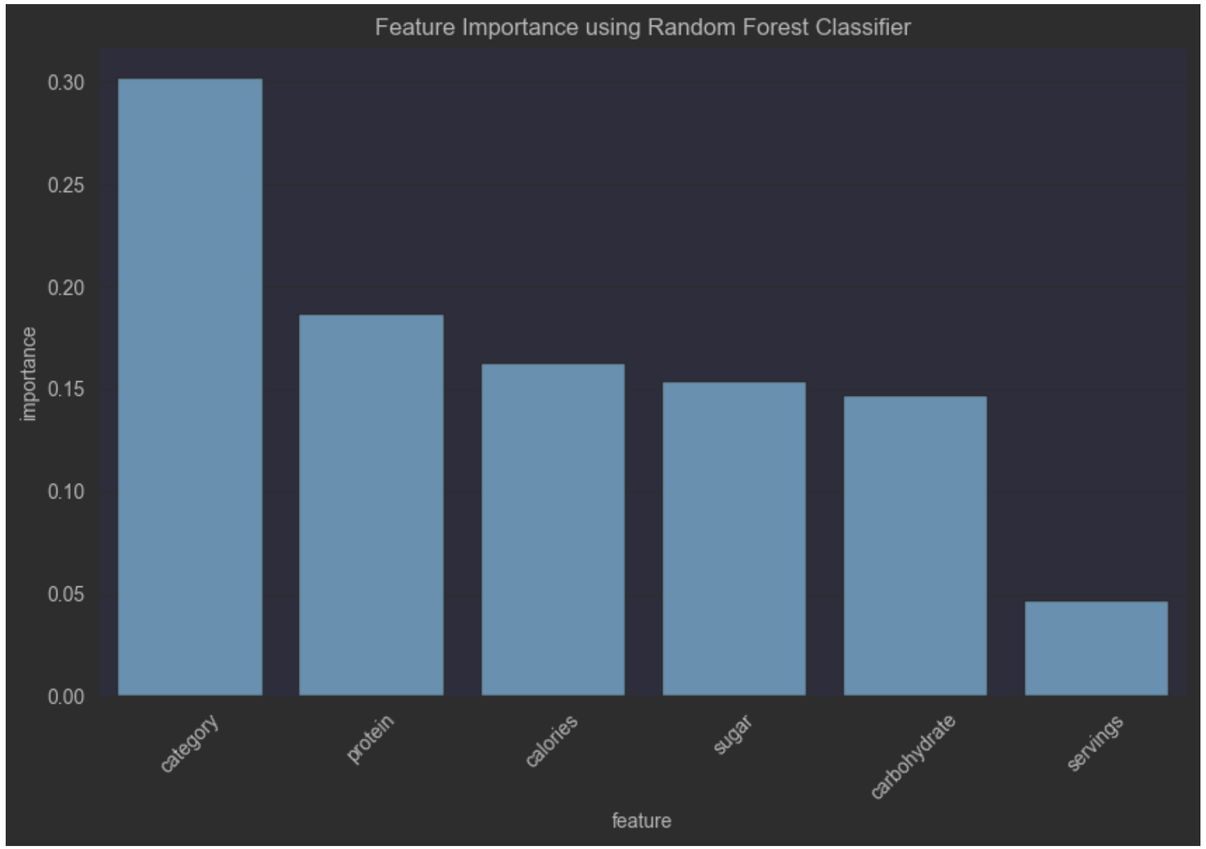


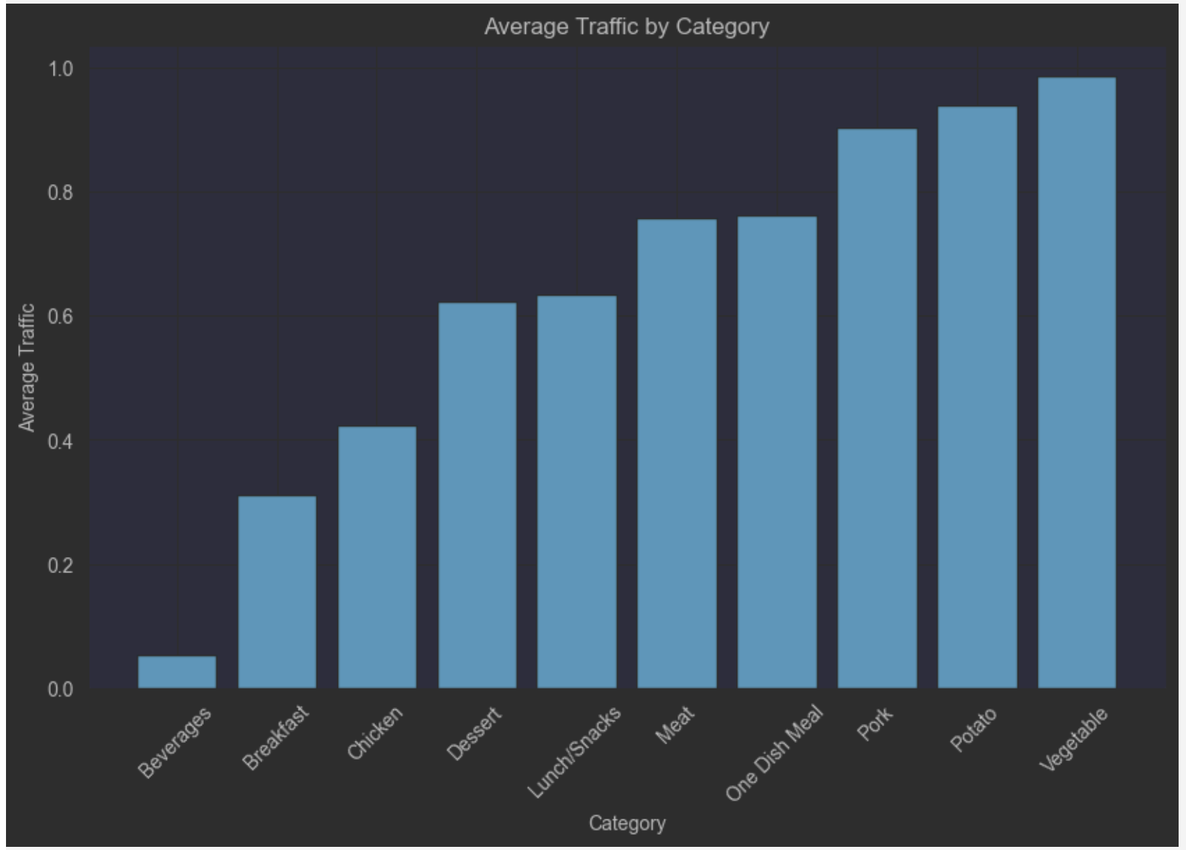
Distribution Similarity / Low Correlation Coefficient

Variables with similar distributions between high_traffic categories (true vs. false) may contribute less to predictive modeling. Their similar distributions suggest they might not significantly enhance the model's ability to predict high_traffic status. Additionally, weak correlations between these numerical variables (calories, carbohydrates, protein) and high_traffic further indicate that they may not effectively distinguish between categories. Despite this, we will retain these features to provide additional dimensions, interaction effects for our model.

Important Features

The result shows that the category column shows the most importance in predicting the high_traffic column. This is consistent with the observation that the category column has a significant impact on the traffic. This is also backed up by the correlation matrix of the numerical variables and high_traffic.



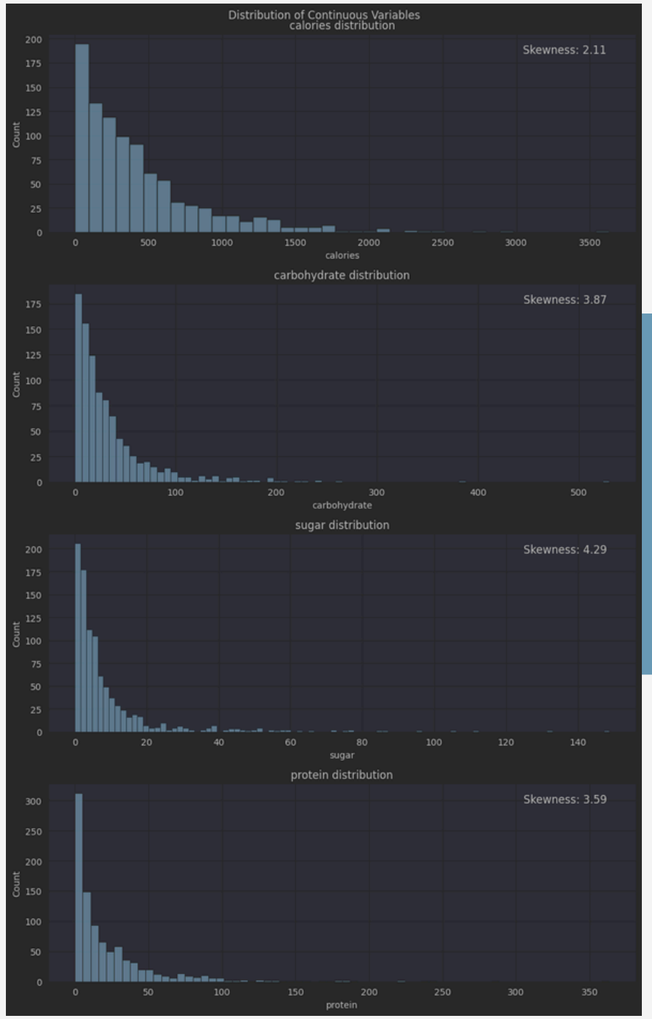


Building the Models

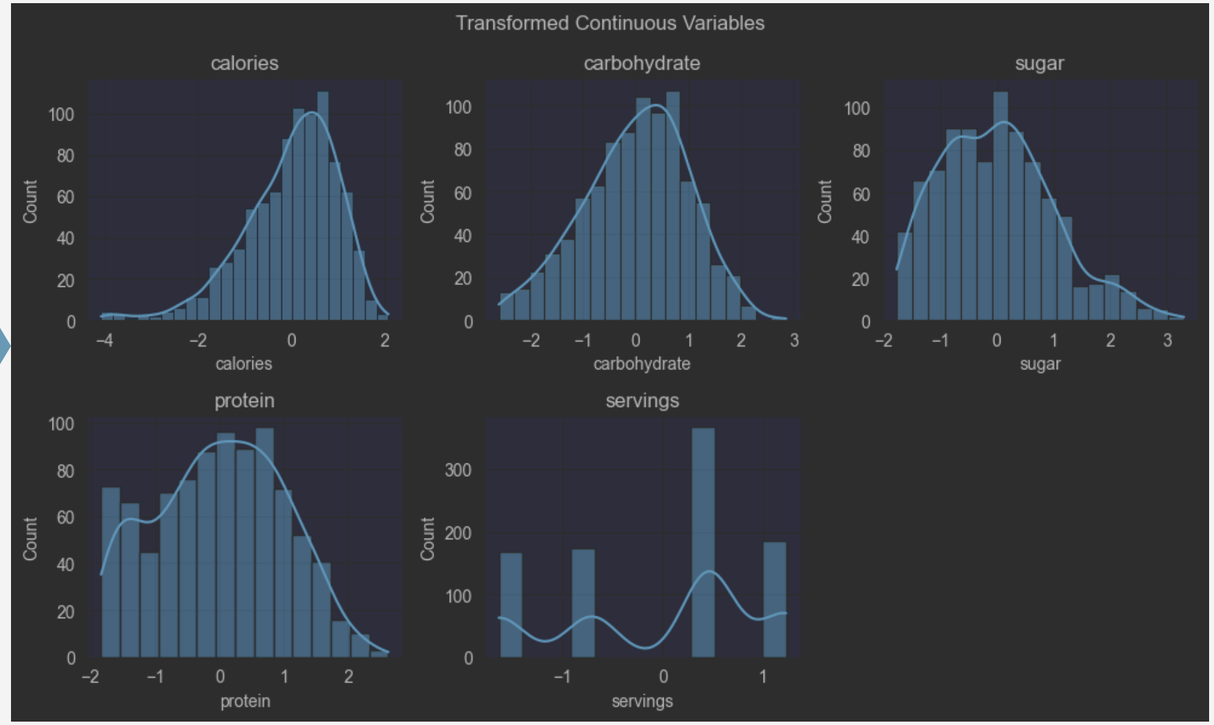
Since the target variable, high_traffic, is binary (True if traffic is High, false otherwise), then this is a classification problem. We use the logistic regression and XGBoost models.

Logistic Regression (Baseline Model)	Uses the features Category (One Hot Encoded) and Servings. A baseline model, where we can compare other models based on how this model performs
Logistic Regression (Fine Tuned)	Uses the features : category (One-Hot Encoded) and all the numerical features -> logarithmical transformations for numerical features except servings -> Standardization using StandardScaler.
XGBoost (comparison model)	Uses the same features as Logistic Regression with Standardization. Used Hyperparameter tuning to find the best hyperparameters.

The need for data transformation and standard scaling



Applying logarithmic transformation and standardize features by removing the mean and scaling to unit variance.



RESULTS

The models are trained using 80% of the data size, while it is tested using the remaining 20% for evaluation.

Logistic Regression (Baseline)

Results for Base Model:

Accuracy: 0.80, Precision: 0.84, Recall: 0.84, F1 Score: 0.84

	precision	recall	f1-score	support
False	0.74	0.75	0.75	69
True	0.84	0.84	0.84	110
accuracy			0.80	179
macro avg	0.79	0.79	0.79	179
weighted avg	0.81	0.80	0.80	179

Logistic Regression (Fine Tuned)

Results for Standardized Logistic Regression:

Accuracy: 0.81, Precision: 0.86, Recall: 0.85, F1 Score: 0.85

	precision	recall	f1-score	support
False	0.72	0.74	0.73	62
True	0.86	0.85	0.85	117
accuracy			0.81	179
macro avg	0.79	0.79	0.79	179
weighted avg	0.81	0.81	0.81	179

XG Boost (Comparison)

Results for XGBoost Model (Comparison Model):

Accuracy: 0.75, Precision: 0.80, Recall: 0.77, F1 Score: 0.78

	precision	recall	f1-score	support
False	0.70	0.73	0.71	75
True	0.80	0.77	0.78	104
accuracy			0.75	179
macro avg	0.75	0.75	0.75	179
weighted avg	0.76	0.75	0.75	179

Best Model: Logistic Regression (Fine Tuned)

All models (Base Logistic Regression, Standardized Logistic Regression, and XGBoost) achieved an accuracy of 75% to 81%, with similar performance across key metrics. The models show a strong balance between precision and recall, indicating effective identification of high-traffic recipes. The **recall** values are particularly noteworthy, as they highlight the models' **ability to correctly identify most high-traffic recipes**, which is crucial for **maximizing traffic and potential subscriptions.**

Metric for the Business to Monitor

Monitoring Business Objectives

To align with business goals like increasing traffic and subscriptions, focus on:

- **Key Metrics or Key Performance Indicator:**
  Traffic Impact: Changes in website engagement.
  Subscription Growth: Number of new subscriptions.
  Recipe Performance: User interaction with featured recipes.
- **Measurement Methods:** Use web analytics tools for traffic and engagement. Track subscriptions via CRM systems. Analyze recipe performance through clicks and views.
- **Benchmarks:** Traffic: Use historical data or industry averages. Subscriptions: Set goals based on current averages. Recipe Performance: Define success criteria like increased interactions.
- **Regular Monitoring:** Implement a reporting system for weekly or monthly updates. Compare with benchmarks to assess progress.
- **Actionable Insights:** Adjust strategies based on metrics to optimize recipe selection and improve engagement.

Estimating Initial Metric Values

- **Website Traffic:**
  - Current Average: 10,000 visits per week.
  - Baseline Value: This will be used to gauge the impact of featuring popular recipes.
- **Subscription Rates:**
  - Current Average: 200 new subscriptions per month.
  - Baseline Value: This figure serves as the starting point for measuring the effectiveness of increased traffic on subscription growth.
- **User Engagement:**
  - Current Metrics: Average time spent on site and pages viewed per visit.
  - Baseline Value: To assess changes in user engagement following model implementation.
- **Projected Improvements:** Based on the model's performance and observed impact from similar initiatives, a projected increase of 20% in website traffic and a proportional rise in subscription rates are anticipated.
- By defining and monitoring these metrics, the business can systematically track the success of model-driven changes and make informed decisions to optimize recipe selection for increased traffic and subscriptions.

Why the business should use the model?

1

Optimizing Homepage Recipes

- Objective: Predict which recipes will lead to high traffic when displayed on the homepage.
- Benefit: Data-driven decision-making enhances the likelihood of attracting more visitors.

2

Increasing Website Traffic

- Impact: More traffic is directly linked to higher engagement and potential revenue.
- Observation: Traffic can increase by up to 40% when popular recipes are featured.

3

Boosting Subscriptions

- Connection: Higher traffic often results in increased subscriptions.
- Outcome: Displaying popular recipes can improve conversion rates and contribute to company growth.

4

Improving Decision-Making

- Method: Provides a systematic, data-driven approach to recipe selection.
- Advantage: Replaces subjective choices with objective, analytical decisions based on historical data.

5

Enhancing Customer Experience

- User Benefit: Features recipes likely to be popular, improving visitor satisfaction.
- Engagement: Better user experience through relevant and interesting content.

6

Strategic Resource Allocation

- Efficiency: Insights from the model allow for better allocation of marketing and promotional resources.
- Focus: Promoting popular recipes can lead to better returns on marketing investments.

7

Competitive Advantage

- Edge: Utilizes advanced predictive models to stay ahead of competitors.

Summary

In summary, the development and evaluation of our predictive models have successfully identified the most effective approach for enhancing recipe selection on the Tasty Bytes homepage. By utilizing the Standardized Logistic Regression model, which demonstrated superior accuracy and recall, we can accurately predict high-traffic recipes, thereby driving more visitors to the website and potentially increasing subscriptions.

Our thorough data validation and exploratory analysis have laid a strong foundation for reliable predictions. Although some features showed similar distributions between high-traffic categories, they still contribute valuable dimensions to our model.

By implementing the recommended model and monitoring key metrics such as recipe click-through rates, traffic increases, and subscription rates, Tasty Bytes will be well-positioned to optimize homepage content, engage users more effectively, and achieve its growth objectives. Regular model updates and performance evaluations will further enhance our ability to make data-driven decisions and maximize business impact.