


Predicting High Traffic Recipes

 Felix Odhiambo

 Contact: otienofelix@live.com

 LinkedIn: <https://www.linkedin.com/in/felixoodhiambo/>

Table of Contents

Table of Figures	2
Data Validation and Cleaning.....	3
Exploratory Analysis	3
Target variable	3
Numeric variables	5
Model Fitting and Evaluation	6
Prepare for Data Modelling.....	6
Logistic Regression Model	7
Logistic Regression Evaluation and Assessment	7
Random Forest Model	7
Random Forest Evaluation	7
Results.....	7
Key Findings	7
Business Metrics.....	8
Recommendation	8

Table of Figures

Figure 1: High vs. Low Traffic Recipes.....	4
Figure 2: Calories Distribution.....	4
Figure 3: Relationship between variables during high and low traffic periods.	5
Figure 4: Common Recipes during the High and Low Traffic.	5
Figure 5: High and Low traffic recipes.	6
Figure 6: Logistic Regression Model.....	8

Data Validation and Cleaning

The dataset contains 947 rows and 8 columns before cleaning and validation. I have validated all the columns against the criteria in the dataset table:

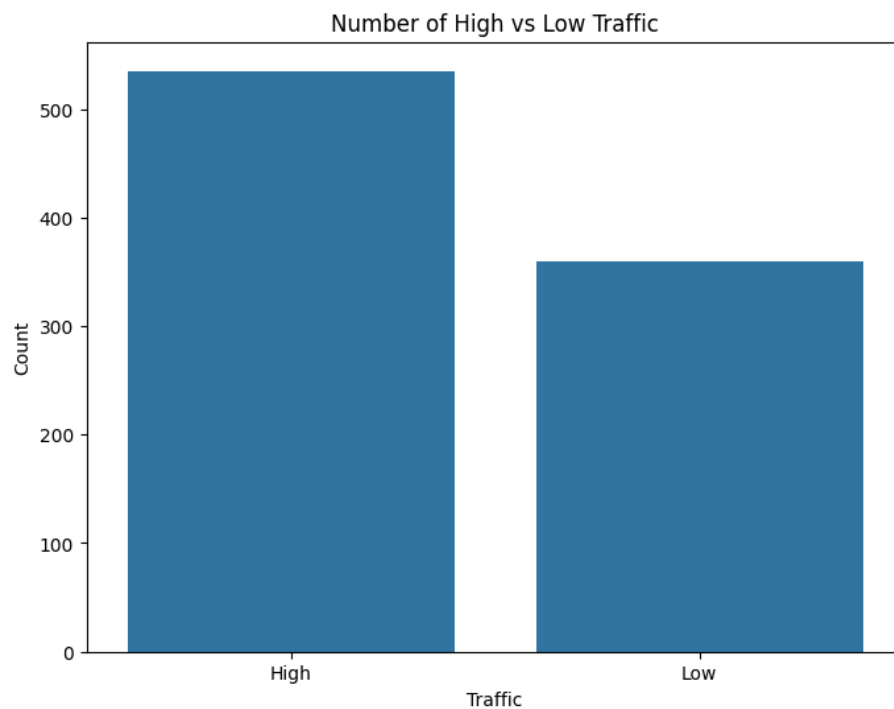
- recipe: Numeric values and unique identifier of the recipe. No cleaning was done.
- calories: Numeric values representing the number of calories. There were 52 missing values which were removed.
- carbohydrate: numeric values representing amount of carbohydrates in grams. There were 52 missing values which were removed.
- sugar: Numeric values representing the amount of sugar in grams. There were 52 missing values which were removed.
- protein: Numeric values representing the amount of protein in grams. There were 52 missing values which were removed.
- category: Character values representing the type of recipe. There were 11 groups instead of the 10 possible groupings. 94 Chicken Breast values were changed to chicken. There were no missing values.
- servings: Numeric and character values representing the number of servings. 4 as a snack and 6 as a snack were converted to numeric values.
- high _ traffic: Character values representing when traffic was high when the recipe was shown. Since traffic can either be high or low, the missing values were replaced with Low. After the data validation, the dataset contains 895 rows and 8 columns.

Exploratory Analysis

Target variable

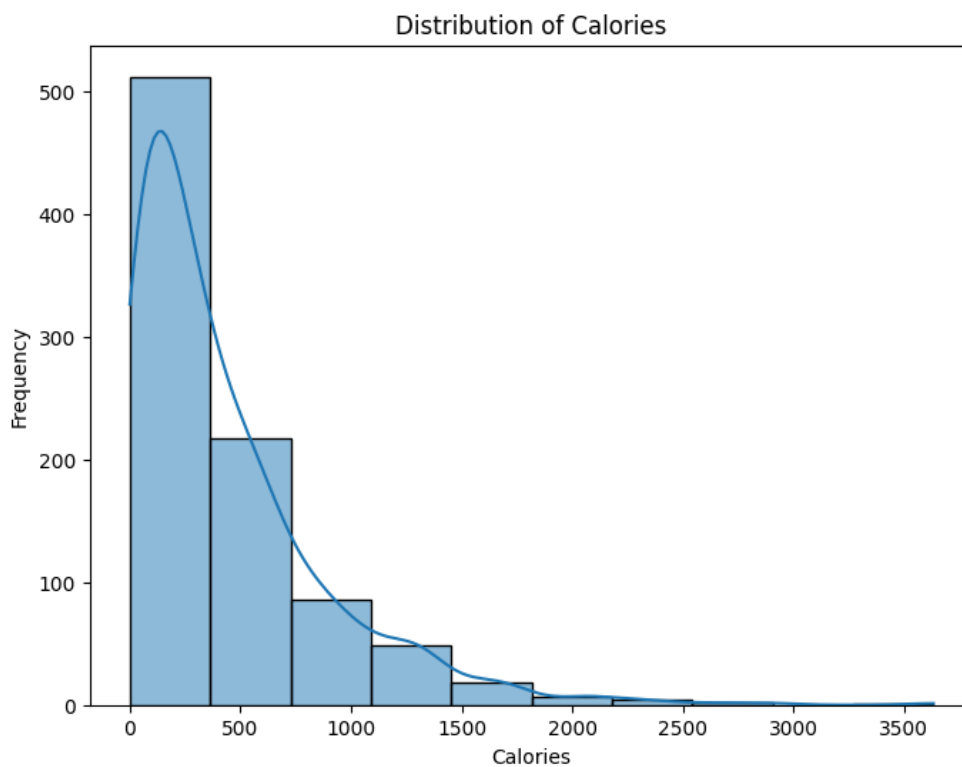
The target variable is high_traffic. This is the variable we want to predict - whether a recipe will lead to high traffic or not. There were over 500 recipes that reported high traffic (Figure 1).

Figure 1: High vs. Low Traffic Recipes



The histogram shows most recipes tend to be lower in calories, with a steep drop off in frequency after 800 calories. The highest frequencies are recipes in the 0-400 and 400-800 calorie ranges. The most shown recipe categories have low calories (Figure 2).

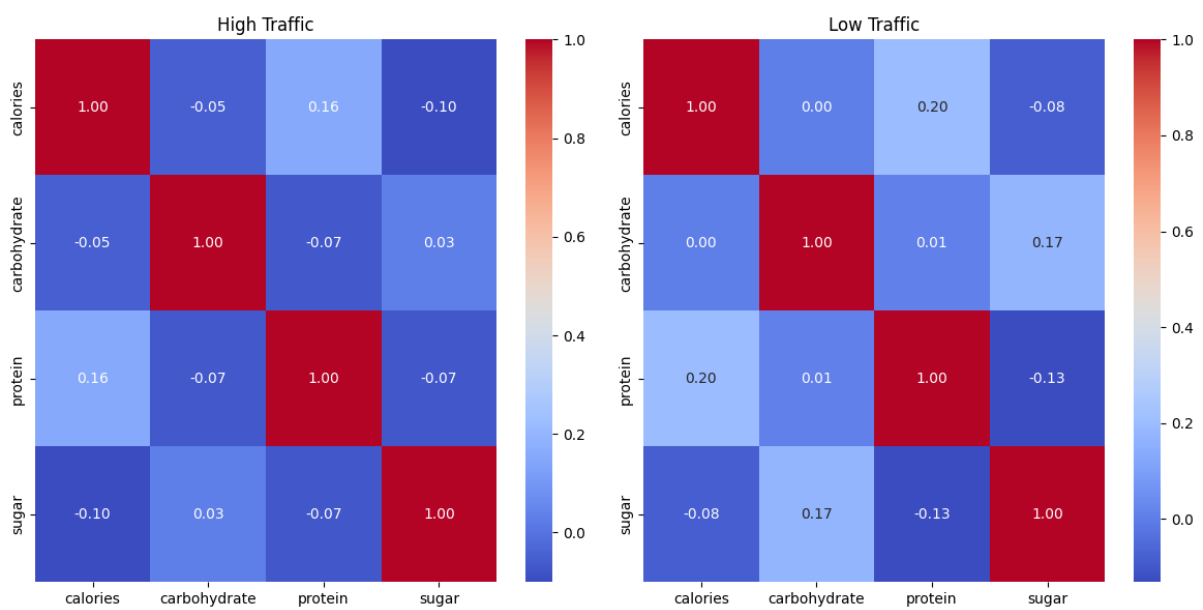
Figure 2: Calories Distribution



Numeric variables

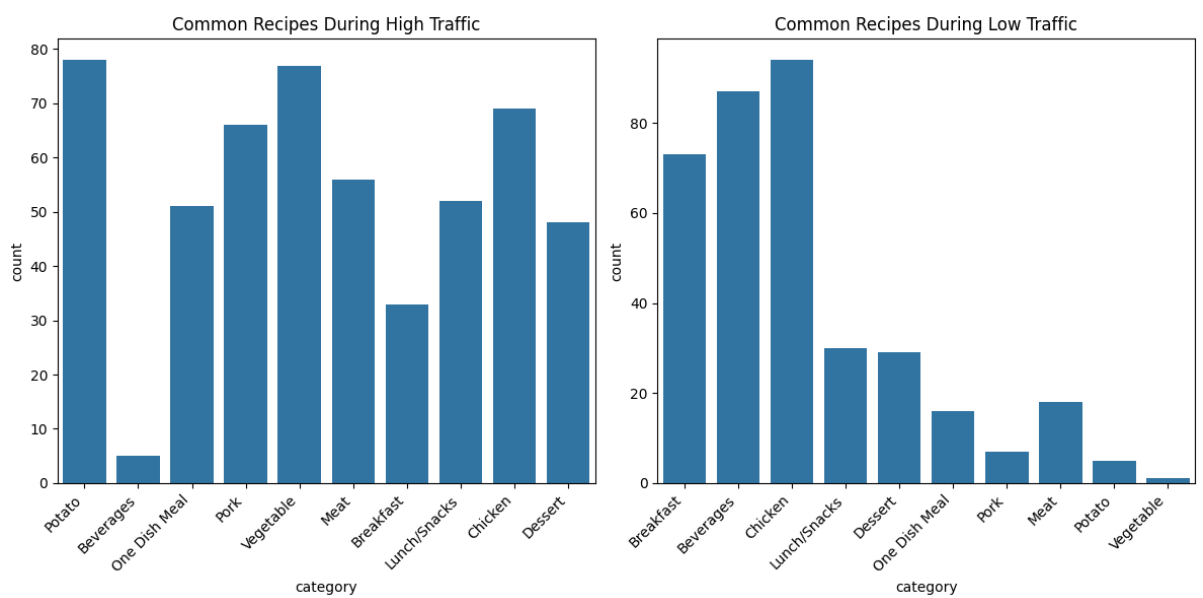
In high traffic there were negative relationships between carbohydrate and calories, sugar and calories, protein and carbohydrate, and protein and sugar, while in low traffic there were negative relationships between sugar and calories and protein and sugar (Figure 3).

Figure 3: Relationship between variables during high and low traffic periods.



The histogram shows that potato was the common recipe during the high traffic while beverages were the least common recipe. On the other hand, Chicken was the most common during the low traffic while vegetable was the least common recipe (Figure 4).

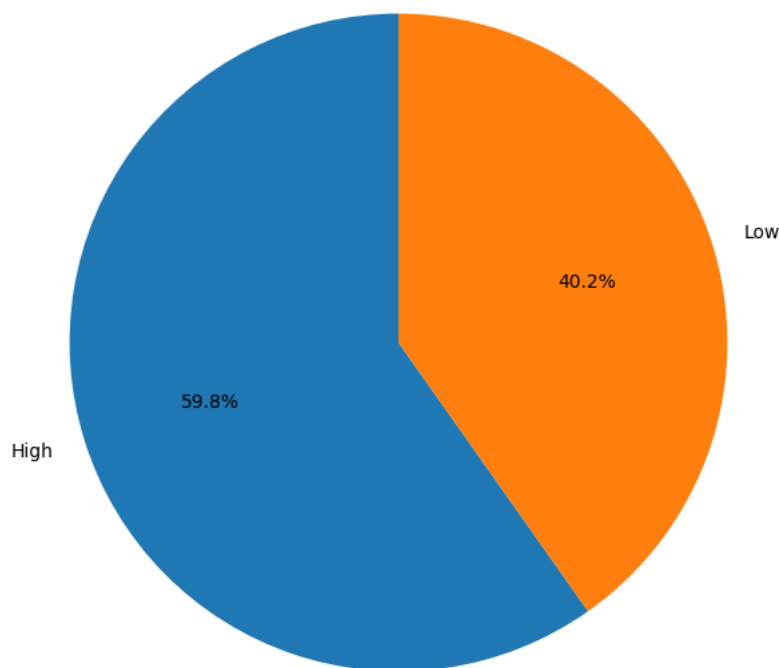
Figure 4: Common Recipes during the High and Low Traffic.



Model Fitting and Evaluation

I am choosing the Logistic Regression model to predict the effects of feature variables (IVs) on target variable. The comparison model I am choosing is Random Forest. I used accuracy, precision, and recall scores to evaluate the model because the dataset was balanced (Figure 5).

Figure 5: High and Low traffic recipes.



Prepare for Data Modelling

To enable modelling, we chose calories, carbohydrate, sugar, protein, category, and servings as features, and high_traffic as target variables. I also have made the following changes:

- Convert the categorical variables into numeric features
- Split the data into a training set and a test set

Logistic Regression Model

Logistic Regression Evaluation and Assessment

Accuracy: 0.7709

Precision: 0.7286

Recall: 0.6986

Random Forest Model

Random Forest Evaluation

Accuracy: 0.7151

Precision: 0.6571

Recall: 0.6301

Results

The accuracy, precision, and recall scores of the Logistic regression were 0.7709, 0.7286, and 0.6986 respectively, while that of Random Forest were 0.7151, 0.6571, and 0.6301. The higher accuracy, precision, and recall scores means that the logistic regression model fits the feature variables better in predicting the target variable than Random Forest model.

Key Findings

- About 24.16% of the variance in high traffic can be explained by the model.
- Sugar and Servings significantly increase the likelihood of high traffic.
- Category has a strong negative impact on traffic—certain categories attract lower traffic.
- Calories, Carbohydrates, and Protein are not significant predictors (Figure 6).

Figure 6: Logistic Regression Model

```
Optimization terminated successfully.
Current function value: 0.510667
Iterations 6
```

Logit Regression Results						
Dep. Variable:	high_traffic	No. Observations:	716			
Model:	Logit	Df Residuals:	710			
Method:	MLE	Df Model:	5			
Date:	Mon, 10 Mar 2025	Pseudo R-squ.:	0.2416			
Time:	08:50:30	Log-Likelihood:	-365.64			
converged:	True	LL-Null:	-482.12			
Covariance Type:	nonrobust	LLR p-value:	2.479e-48			
	coef	std err	z	P> z	[0.025	0.975]
calories	0.0004	0.000	1.843	0.065	-2.27e-05	0.001
carbohydrate	0.0007	0.002	0.355	0.723	-0.003	0.005
sugar	0.0111	0.005	2.103	0.035	0.001	0.022
protein	0.0013	0.003	0.513	0.608	-0.004	0.006
category	-0.4691	0.040	-11.635	0.000	-0.548	-0.390
servings	0.2162	0.039	5.530	0.000	0.140	0.293

Business Metrics

Since the goal of the business is to correctly predict high traffic recipes 80% of the time, we can use recall since it misses 30.14% of high-traffic recipes. Additionally, the business can aim at optimizing their recipe.

Recommendation

To improve recall, include additional features, such as cooking time, ingredients, and cuisine type.

For recipe optimization:

- Increase sugar content (without compromising on health) to gain more traffic.
- Increase servings per recipe to generate more traffic.
- Optimize recipe categories by encouraging the recipes that generate more traffic.