Data Science Professional Exam DS601P

Prediction of ‘High Traffic’ attracting recipe

Prepared by - Zaw Lynn Htut (Mr.)

2024 Aug 25th

Overview

Business Goal

Optimize recipe recommendations to drive user satisfaction and retention.

Project Objective

- Provide data-driven insight on optimization of recipe selection
- Develop an accurate predictive model with 80% accuracy

Tackling Business Issue

1. Find out what makes recipe popular and attract traffics
    a. Data Cleaning
    b. Exploratory Analysis

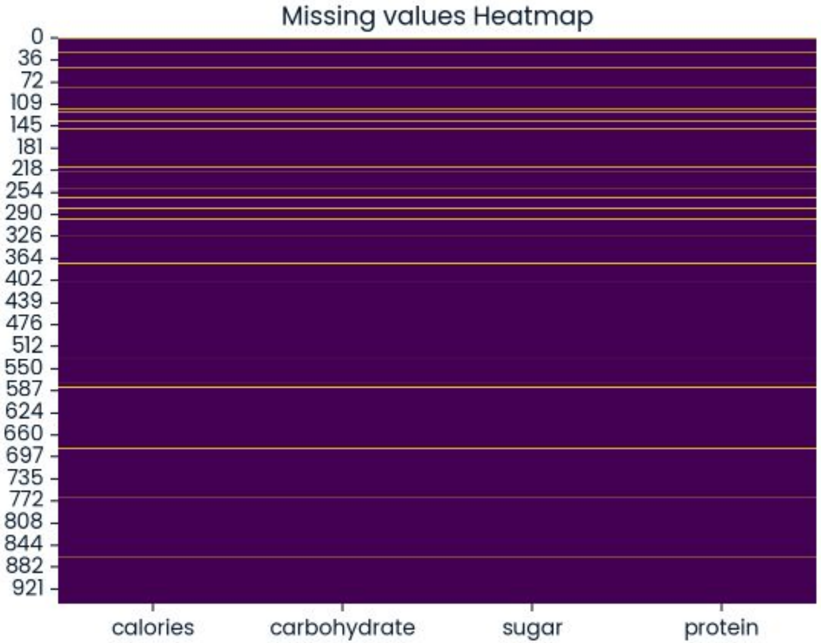
2. Training predictive model with 80% accuracy
    a. Preprocessing
    b. Selection of algorithms
    c. Hyperparameter fine tuning
    d. Model improvement

1. Data Cleaning (Sanitization)

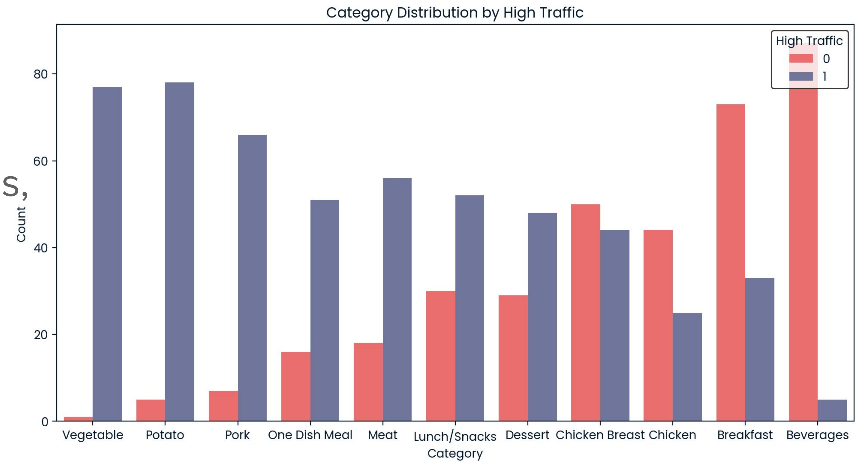
“Accurate Input, accurate output”

To achieve accurate actionable insights:

- Checking for duplicate & remove it
  - To Reduce artificial weight & generalization
- Transformed to most suitable data types (object ⇒ category)
  - Interpretability, Algo compatibility, and model performance
- Handled missing values in numeric columns
  - From missingness heatmap, it’s evident if a value is missing in one column, it's often missing in the other columns for the same row. This pattern suggests a strong correlation between above four columns. Total number of missing value rows - 29 (only 3% of total rows)
  - For Label ‘high_traffic’ column, either ‘high’ or null value. Convert “high” to 1 and null to 0 for better data visualization.

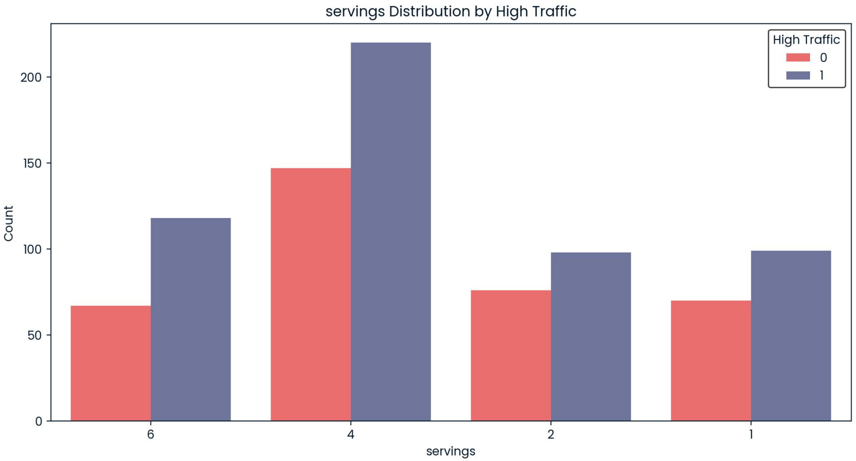


2. Exploratory Analysis



Correlation between Recipe category and traffic:

- It is evident that healthier diet, such as vegetables, is the choice among visitors.



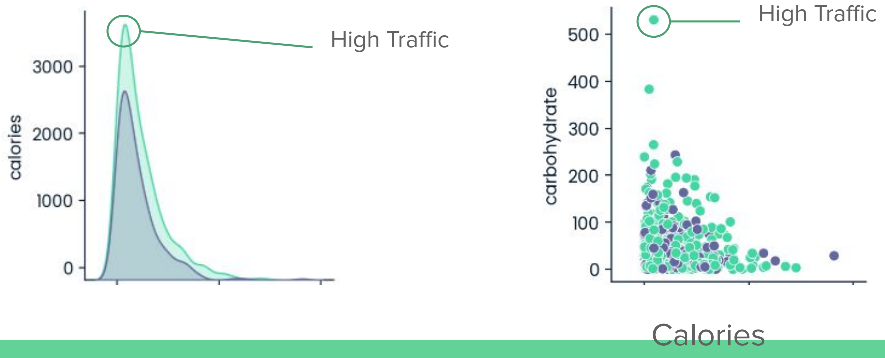
Serving size and traffic:

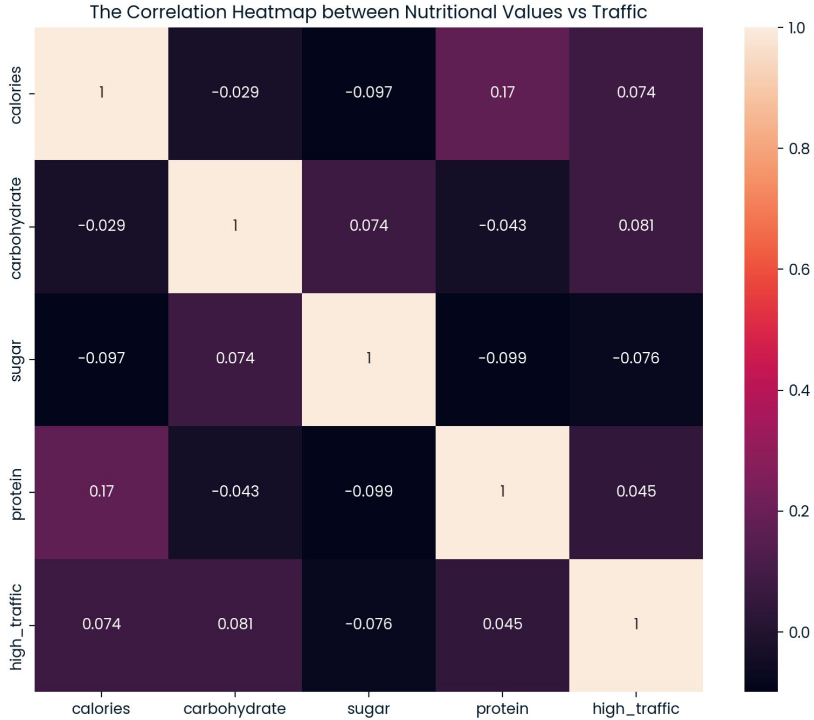
- Size of 4 - 6 appears to be favourite among visitors

3. Exploratory Analysis (Nutritional values)

From the correlation matrix heatmap, it appears that there's an inverse relationship between calories and traffic.
However, histograms and boxplots show that certain low-calorie recipes don't necessarily attract high traffic.

Scatterplot shows relationship between calories, carbs, and traffic.
High-carb, high-calorie content isn't the sole driver of popularity.
Calorie and Carbohydrate Content: Important predictors of high-traffic recipes.




The Correlation Heatmap between Nutritional Values vs Traffic

	calories	carbohydrate	sugar	protein	high_traffic
calories	1	-0.029	-0.097	0.17	0.074
carbohydrate	-0.029	1	0.074	-0.043	0.081
sugar	-0.097	0.074	1	-0.099	-0.076
protein	0.17	-0.043	-0.099	1	0.045
high_traffic	0.074	0.081	-0.076	0.045	1

4. Predictive Model Training (without hyperparameter tuning)

Feature Selection : Recipe category plays major role in relation with traffic, follows by nutritional values , and serving size.

Class balancing: Label column high traffic has 60% high and 40% low. Using smote to rebalance.

Encoding & Scaling: label encoding for ‘category’ & ‘servings’ and scaling for nutritional value columns.

Training & Validation : 80% training and 20% testing split

Algo choice : Since this is binary classification, choices are Random forest, Logistic regression, XGBoost , and LGBM (for future large dataset training).

Overall contender: Random Forest , recall 0.74 with the highest overall accuracy and F1-score, indicating a strong balance between precision and recall.

4. Predictive Model Training (hyperparameter tuning & feature engineering)

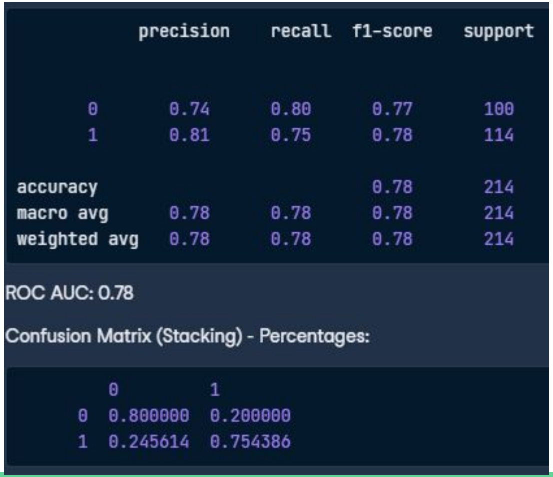
Additional Feature : Since calories and carbs has affinity towards high traffic, diving calories with carbs would results in another important feature.

Gridsearchcv: To improve model performance, incorporating grid search cross validation and this will improve efficiency explorations.

Stacking to try meta-model: to improve generalization and leverages diverse models.

Result : Stacking (combined only good traits of all four models)

Stacking's Superiority: The Stacking model achieves the highest overall accuracy and F1-score, confirming its effectiveness in combining multiple models. Feature Importance: The category feature remains the most influential predictor, followed by nutritional factors like protein, calories, and calories_per_carbohydrate. Class-Specific Performance: The model demonstrates a strong ability to correctly predict '1' instances (high-traffic recipes), with a recall score close to or above 0.8, meeting the desired objective.



```
              precision    recall  f1-score   support


           0       0.74      0.80      0.77       100
           1       0.81      0.75      0.78       114

    accuracy                           0.78       214
   macro avg       0.78      0.78      0.78       214
weighted avg       0.78      0.78      0.78       214
```

ROC AUC: 0.78

Confusion Matrix (Stacking) - Percentages:

```
          0         1
0  0.800000  0.200000
1  0.245614  0.754386
```

5. Conclusion & Final Recommendations

Most Important Features:

-	Recipe category	(40%)
-	Protein	(14%)
-	Calories	(11%)
-	Calories per carbs	(10%)
-	Sugar	(10%)

Result : Stacking (meta-model) best overall

Although Model results didn't achieved 80% mark, the Stacking model achieves a recall of 0.75 for class '1', which is close to the desired 80% threshold. This indicates that it is effectively predicting high-traffic recipes.

Recommendation: larger dataset, less missing values, and time of traffic