Predicting Popular Recipes

Tasty Bytes

Presented to the Product Manager

Overview

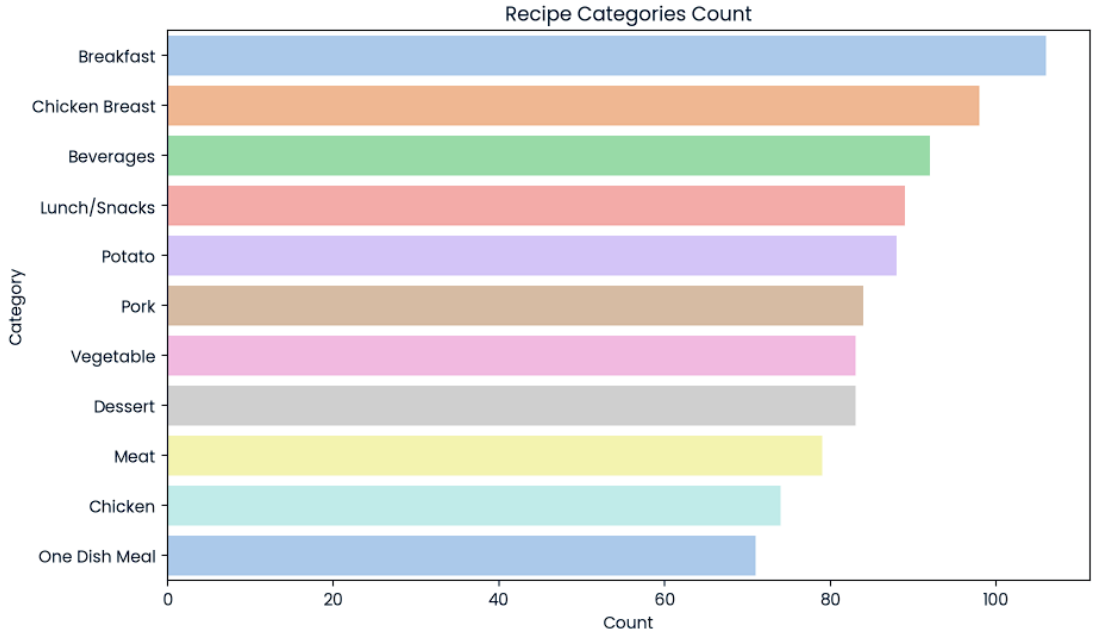
- High-traffic recipes increase overall website traffic by up to 40%, leading to more subscriptions
- Predict recipes that will lead to high traffic on the website
- Minimize the chance of displaying unpopular recipes
- **A solution to correctly predict high-traffic recipes 80% of the time**

Dataset

recipe	calories	carbohydrate	sugar	protein	category	servings	high_traffic
1					Pork	6	High
2	35.48	38.56	0.66	0.92	Potato	4	High
3	914.28	42.68	3.09	2.88	Breakfast	1	null
4	97.03	30.56	38.63	0.02	Beverages	4	High
5	27.05	1.85	0.8	0.53	Beverages	4	null

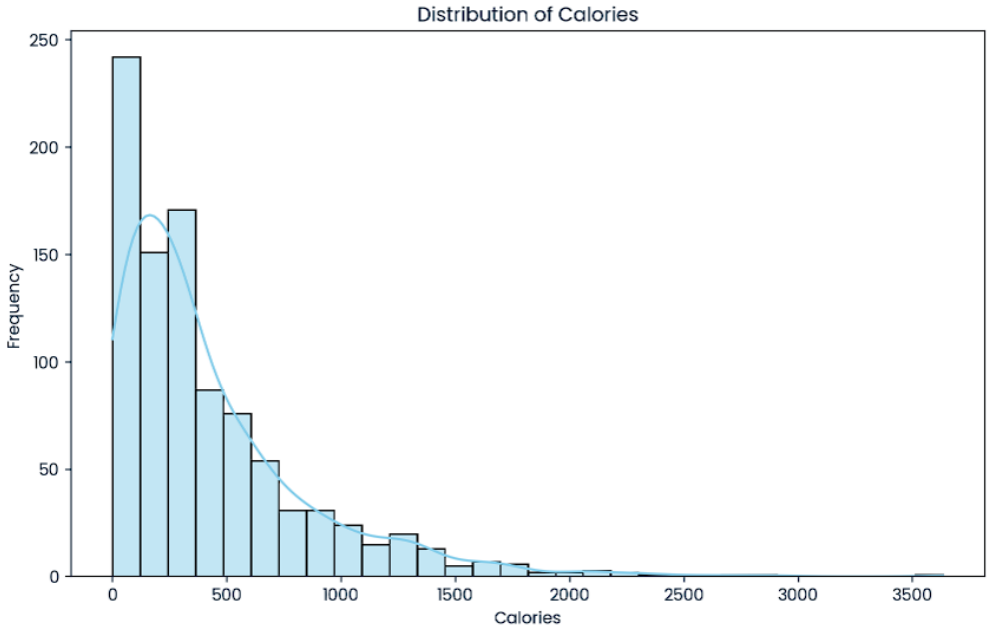
- Data Cleaning:
  - Cleaned inconsistencies in the *servings* column by extracting numeric portion and convert dtype to int.
  - Imputed missing nutritional values with the median value of the respective nutrients.
  - Fill null values under *high_traffic* column with “Low”.
- Exploratory Data Analysis:
  - Checked distribution and variance of variables.
  - Assessed relationships between variables.
- Predictive Modeling:
  - Development of base and comparison models, compared performance metrics and evaluated model performance

Key Findings



Recipe category count

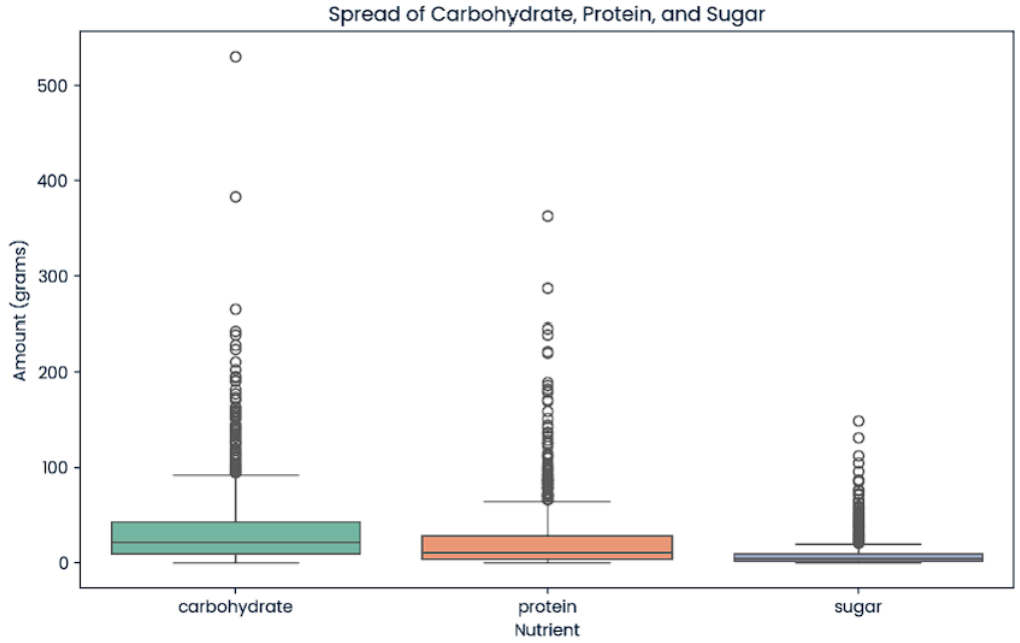
The dataset contains recipes from various categories, with some categories (e.g., "Breakfast“, "Chicken Breast") being more frequent than others. This indicates that certain types of recipes are more common in the dataset.



Distribution of calories

The distribution of calorie is right-skewed, suggesting that most recipes have lower calorie counts, with a few outliers having very high calories. This indicates that most recipes are relatively low-calorie, with a small number of high-calorie recipes.

Key Findings

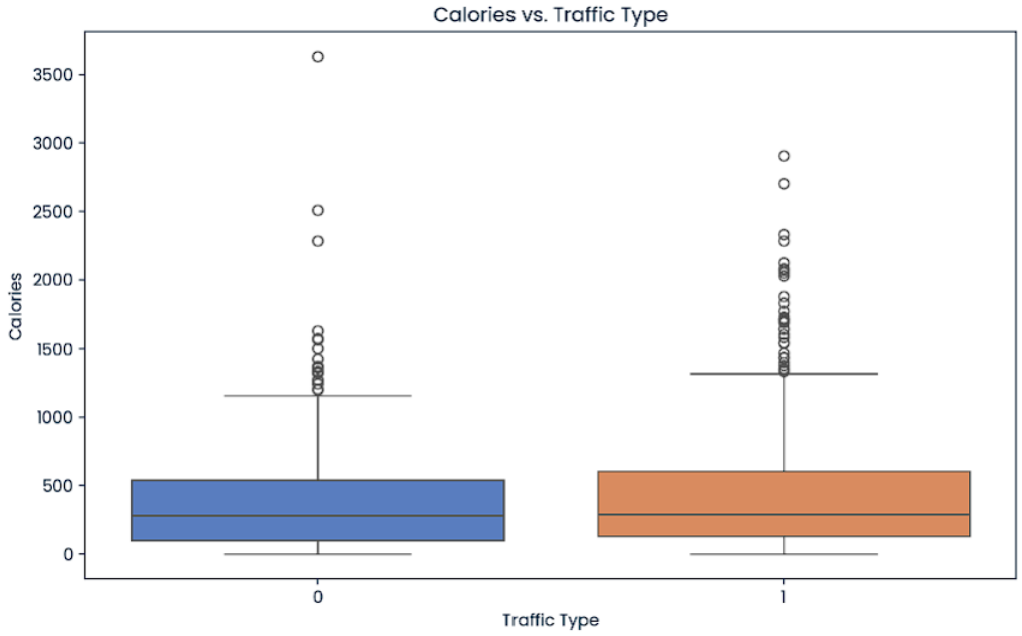


Spread of carbohydrate, protein, sugar

Carbohydrate: The spread is wide, with many outliers on the higher end. This suggests that some recipes are very high in carbohydrates.

Protein: The spread is narrower compared to carbohydrates, with fewer outliers. Most recipes have moderate protein content.

Sugar: The spread is relatively narrow, but there are some outliers with very high sugar content.



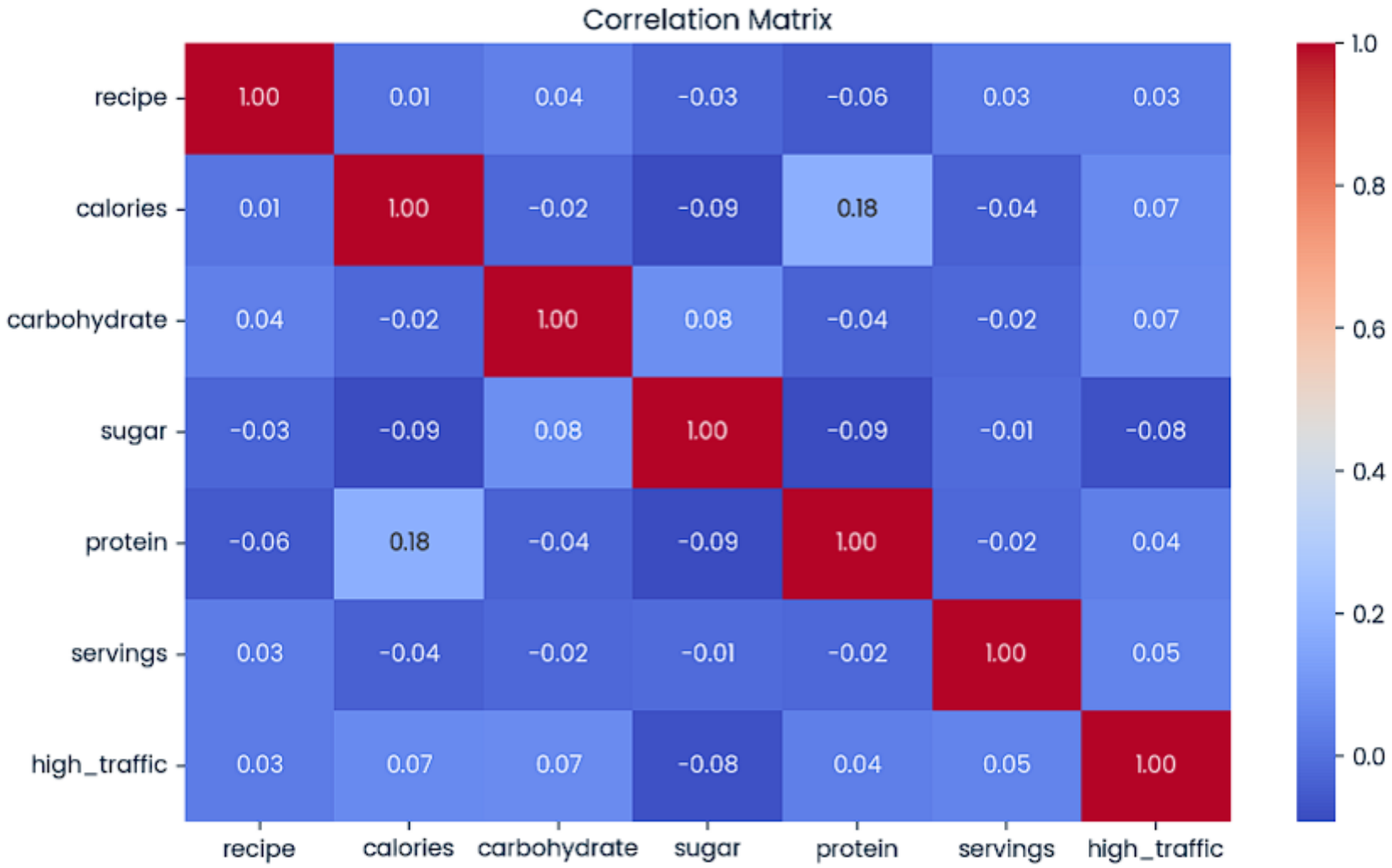
Traffic VS Calorie

"High" traffic recipes tend to have more varied calorie counts, but not dramatically different from "Low".

Key Findings

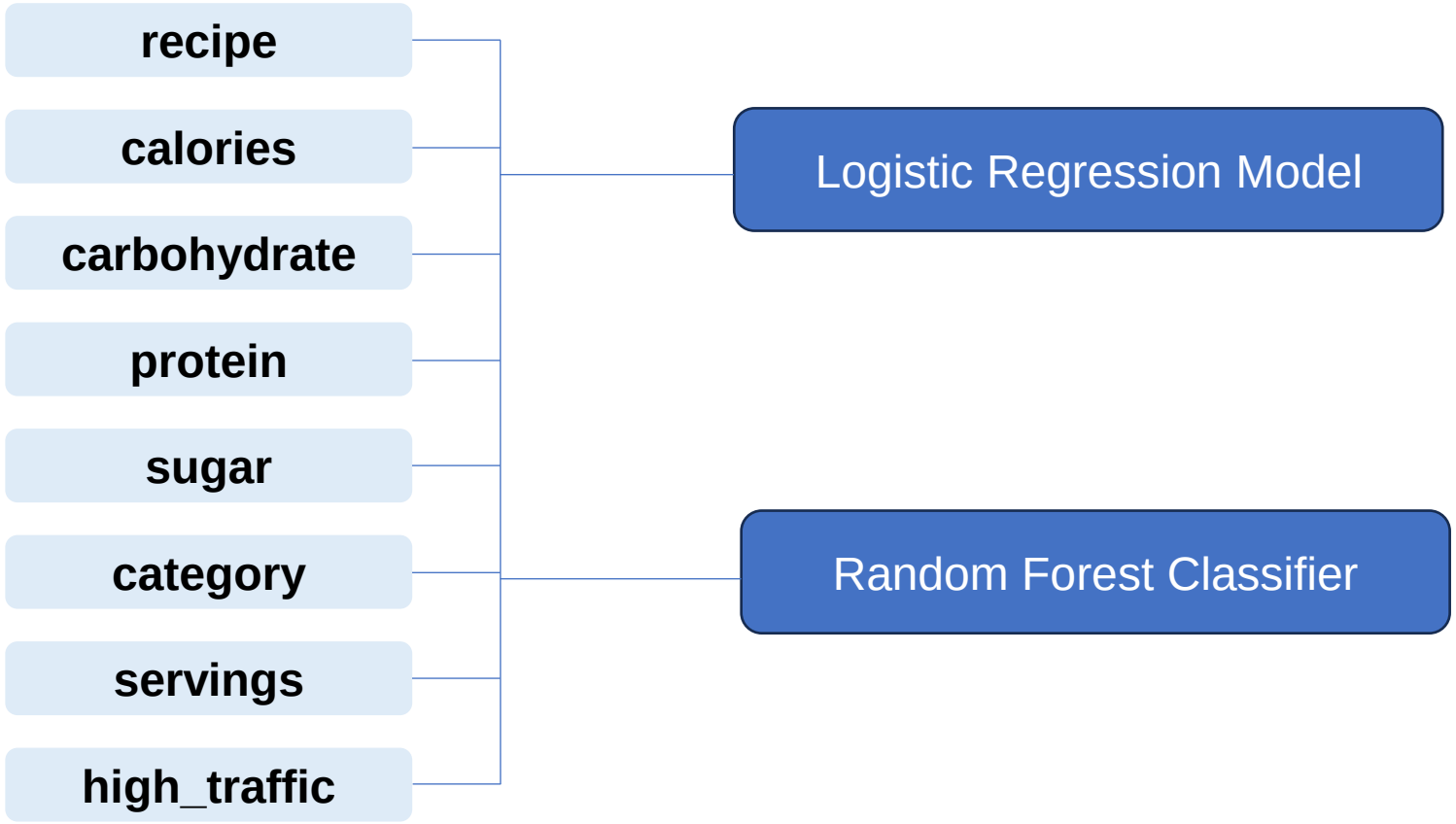
Correlation Matrix

- The heatmap shows correlations between numerical variables (e.g., calories, carbohydrate, sugar, protein).
- There is a moderate positive correlation between calories and protein, as seen in the scatter plot.
- Other correlations (e.g., carbohydrate and sugar) are also visible, indicating relationships between nutritional components.



Model Development

Two models – Logistic Regression model and Random Forest Classifier



Key metrics

ACCURACY

Assess the model's overall detection accuracy across all classes.

Accuracy = TP + TN / (TP + TN + FP+ FN)

PRECISION

Measures the model's accuracy in identifying relevant objects by calculating the ratio of true positives to all detections.

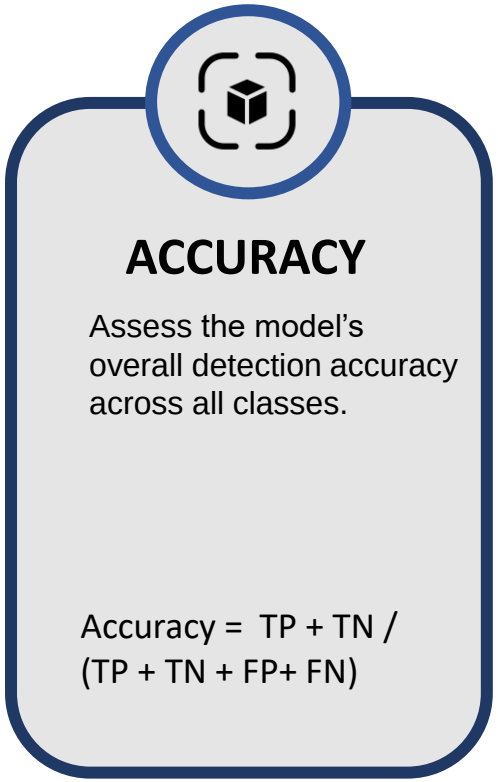
Precision = TP / (TP + FP)

RECALL

Evaluates the model's ability to detect all true objects, indicating the proportion of true positives among all actual truth instances.

Recall = TP / (TP + FN)

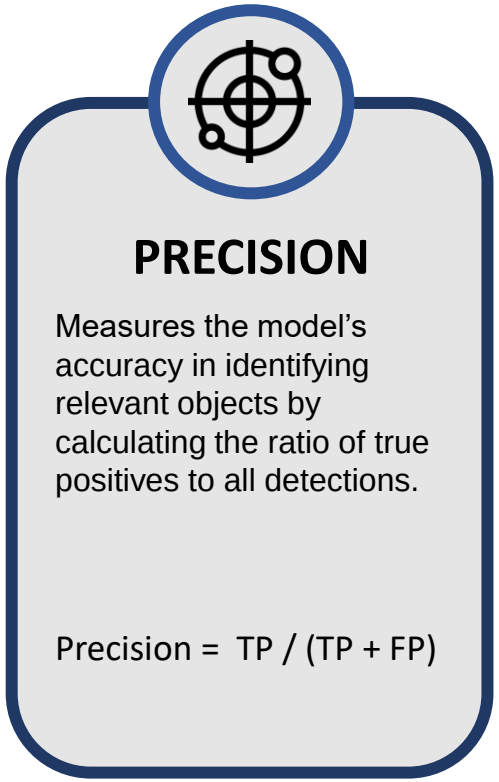
Key metrics



ACCURACY

Assess the model's overall detection accuracy across all classes.

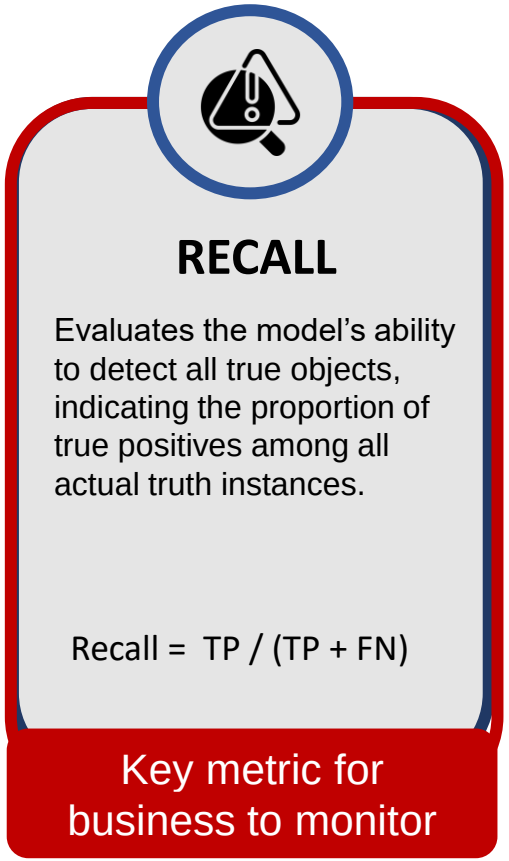
Accuracy = TP + TN / (TP + TN + FP+ FN)



PRECISION

Measures the model's accuracy in identifying relevant objects by calculating the ratio of true positives to all detections.

Precision = TP / (TP + FP)



RECALL

Evaluates the model's ability to detect all true objects, indicating the proportion of true positives among all actual truth instances.

Recall = TP / (TP + FN)

Key metric for business to monitor

Model Evaluation

Model	Accuracy	Precision		Recall	
		Class 0	Class 1	Class 0	Class 1
Logistic Regression	0.75	0.67	0.80	0.67	0.80
Random Forest Classifier	0.70	0.61	0.76	0.59	0.77

1. Precision - Logistic Regression has higher precision for both classes, meaning it makes fewer false positive predictions compared to Random Forest.

2. Recall - Logistic Regression has a higher recall for Class 1, suggesting it correctly identifies more instances of Class 1 compared to Random Forest.

3. Accuracy - Logistic Regression has higher accuracy, meaning it correctly predicts more instances overall compared to Random Forest.

Model Evaluation

Model	Accuracy	Precision		Recall	
		Class 0	Class 1	Class 0	Class 1
Logistic Regression	0.75	0.67	0.80	0.67	0.80
Random Forest Classifier	0.70	0.61	0.76	0.59	0.77

Note: Class 0 – Low-traffic recipes, Class 1 – High-traffic recipes

1. Precision - Logistic Regression has higher precision for both classes, meaning it makes fewer false positive predictions compared to Random Forest.

2. Recall - Logistic Regression has a higher recall for Class 1, suggesting it correctly identifies more instances of Class 1 compared to Random Forest.

3. Accuracy - Logistic Regression has higher accuracy, meaning it correctly predicts more instances overall compared to Random Forest.

Recommendations

- Adopt Logistic Regression as the primary model as it outperforms the Random Forest in terms of precision, recall and accuracy.

- It is more effective in predicting high-traffic recipes, and also has fewer misclassifications (false positives and false negatives), which reduces the risk of incorrect predictions.

- Continuously monitor model performance over time to ensure it remains effective as new data comes in. Track key metrics and detect any degradation in performance.

- Explore feature engineering for further improvements to the model. Test other models to ensure no better alternative exists.

- Since model achieves 80% recall for high-traffic recipes (Class 1), the business can :
    - Prioritize marketing efforts for high-traffic recipes.
    - Optimize inventory management for ingredients used in these recipes.
    - Improve customer satisfaction by promoting popular recipes.

Thank You