

Recipe Site Traffic Prediction

Author: Son Hai Le



- 1. Project Overview**
- 2. Method**
- 3. Key Findings**
- 4. Summary & Recommendations**

1. Project Overview

- Background
- Business Goals

10 MINUTES PREP
30 MINUTES COOK

1 FORK
SERVES 6

SALMON, POTATO, AND PESTO TART

You can use puff pastry from a block. Roll out on a lightly floured surface to a rectangle measuring about 10 x 15 in (25 x 35 cm).

EQUIPMENT

- Saucepan • Colander
- Large, nonstick baking sheet • Sharp knife • Fork
- Mixing bowl
- Wooden spoon
- Rolling pin

INGREDIENTS

- 7 1/2 (250 g) new potatoes
- 13 oz (350 g) frozen puff pastry sheet
- 1 cup (250 g) ricotta cheese
- 2 eggs, beaten
- 3 tbsp fresh pesto, plus extra for drizzling
- 2 tsp grated lemon zest
- 7 1/2 (225 g) cooked salmon, broken into pieces
- Salt and freshly ground black pepper

1 Preheat the oven to 400°F (200°C). Scrub the new potatoes, then thinly slice each one.

2 Boil the potatoes in a pot of lightly salted water for 5–6 minutes, until just tender. Drain and allow to cool slightly.

3 Place the rolled-out puff pastry on a baking sheet. Using a sharp knife, score a 1-in (2.5-cm) rim along the sides of the rectangle, being careful not to cut all the way through. Prick the inside area lightly with a fork.

4 In a bowl, beat together the ricotta cheese, eggs, pesto, and lemon zest with a wooden spoon, until well mixed. Season with a little salt and freshly ground black pepper. Stir in half of the salmon.

5 Spoon the mixture inside the marked edge and tuck the potatoes and remaining salmon over the top. Bake in the center rack of the oven for 25 minutes, until the pastry is risen and the filling cooked.

6 Remove from the oven and serve drizzled with extra pesto.

Great party food!

The rim left uncovered becomes tasty and crisp in the oven.

Fresh pesto

1. Project Overview

→ Background

- Homepage recipes can boost total site traffic by up to 40% if popular.
- Recipe currently selected based on personal preference.
- Increased traffic = more subscriptions → high business impact.

→ Business Goals

- Predict which recipes will lead to high traffic.
- Achieve 80+% accuracy in identifying popular recipes.
- Provide data-driven recommendations for homepage selection

2. Method

- Procedure
- Metrics Selection

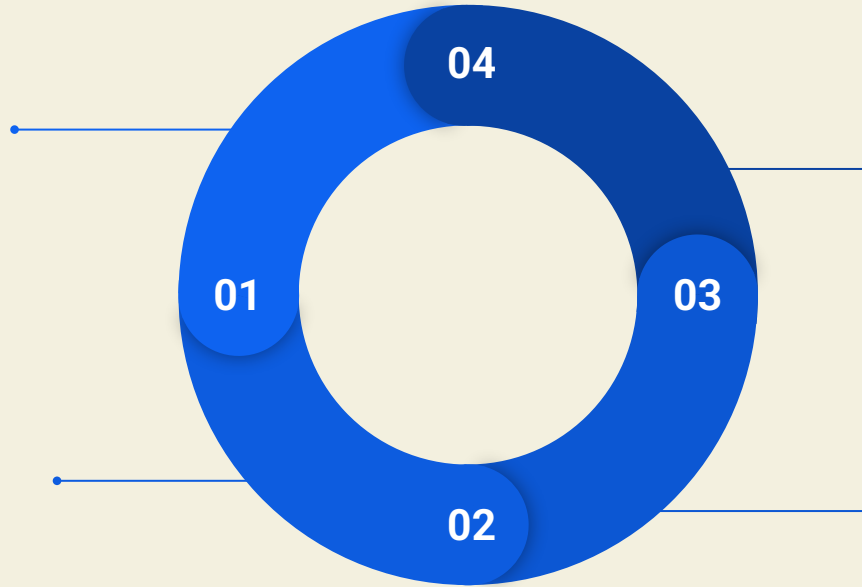
Procedure

Data Validation

1. Validation
2. Cleaning

Exploratory Data Analysis (EDA)

1. Single variable analysis
2. Multiple variable analysis




Model Evaluation

1. Model evaluations on a specific metric


Model Development

1. Data Preprocessing
2. Models Construction and Comparison

Metrics Selection

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


- Overall correctness of the model.
- Can be misleading if popular vs. unpopular recipes are imbalanced.

 **Precision** = $TP / (TP + FP)$

- Of all recipes predicted *popular*, how many actually were?
- High precision = fewer wrongly promoted recipes (i.e., False Positive).

 **Recall** = $TP / (TP + FN)$

- Of all truly *popular* recipes, how many did we catch?
- High recall = fewer missed opportunities.

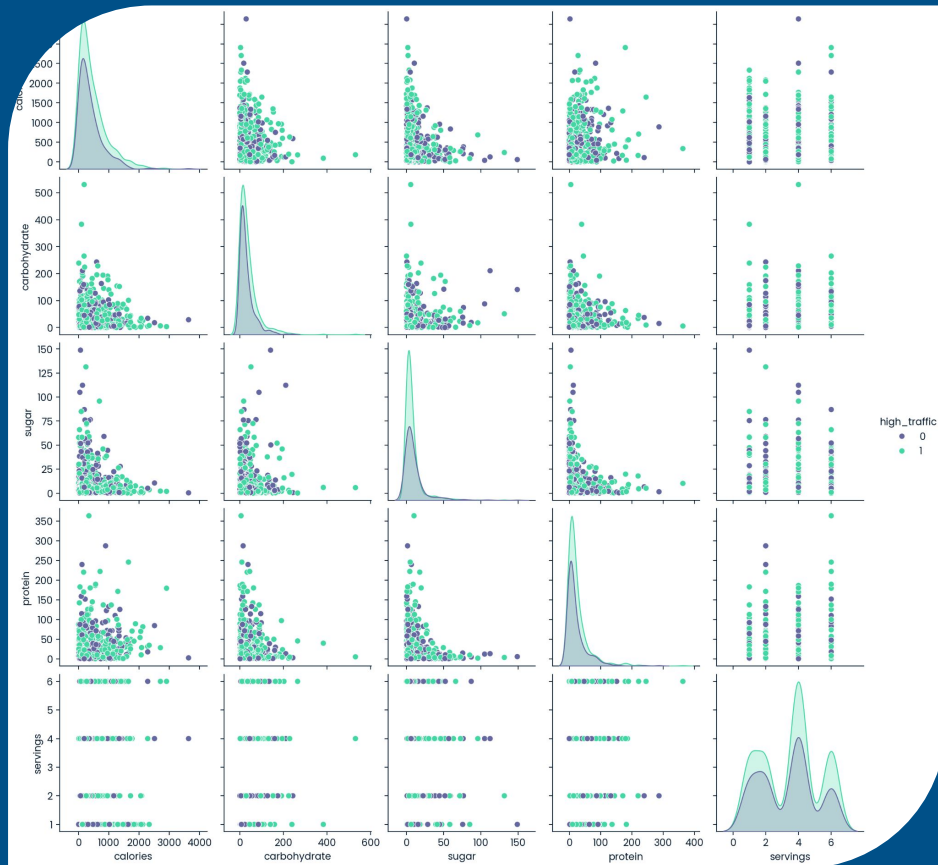
 **F1-Score** = $2 * (precision * recall) / (precision + recall)$

- Balance between precision and recall.
- Useful when having imbalance data.

→ Accuracy and precision: for models comparison
→ F1-score: for hyperparameter tuning

3. Results

- Data Validation
- Exploratory Data Analysis (EDA)
- Data Preprocessing
- Models Construction and Comparison
- Key Findings

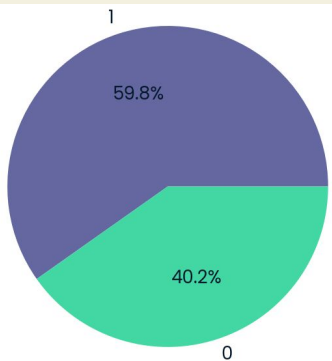


Data Validation

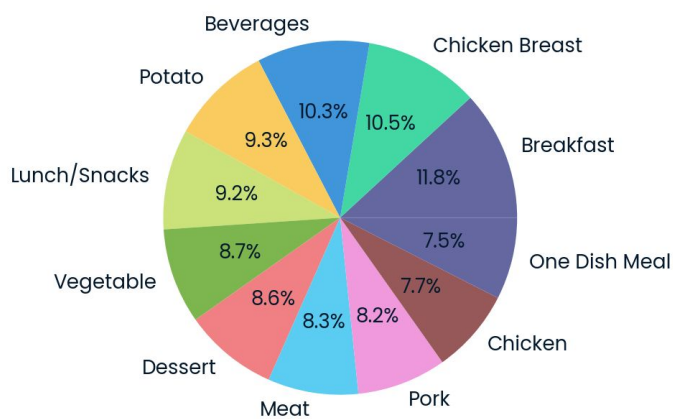
Column	Non-Null Count	Dtype	Clean steps
recipe	947	int64	Remove
calories	895	float64	Remove rows w/ nulls (5.5%)
carbohydrate	895	float64	
sugar	895	float64	
protein	895	float64	
category	947	object	-
servings	947	object	Clean, convert to int
high_traffic	574	object	Replace nulls by 0, High by 1

→ Before cleaning: 947 rows; After cleaning: 895 rows (5.5%↓)

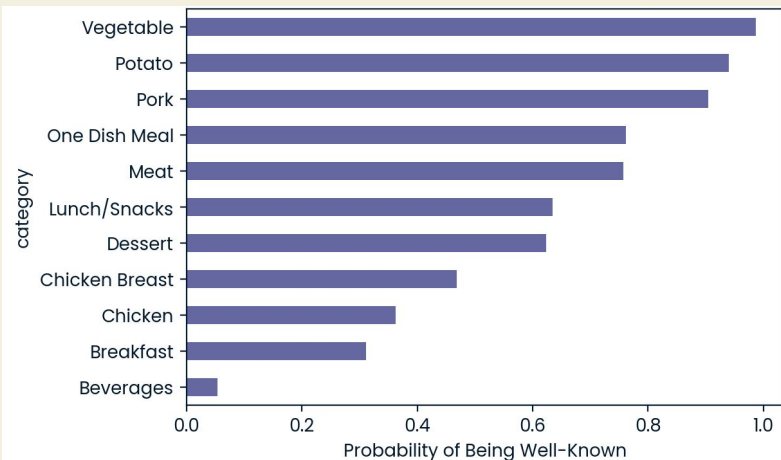
EDA [1/3]



High traffic (1)
/non-high traffic (0)
recipe proportions



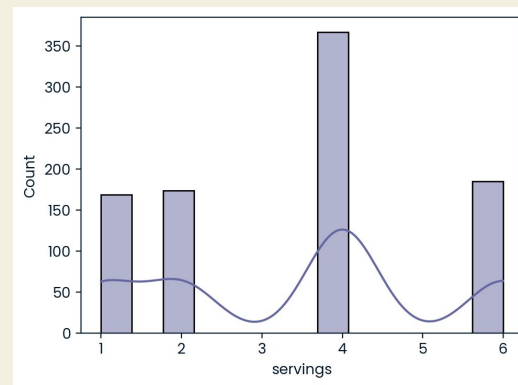
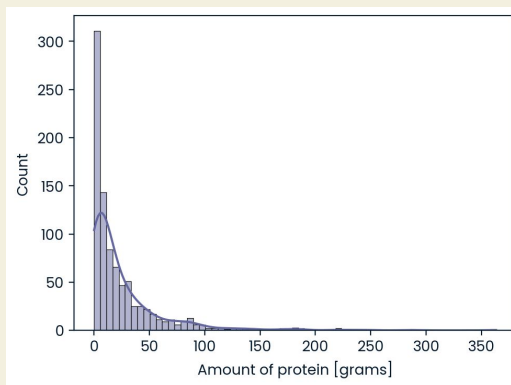
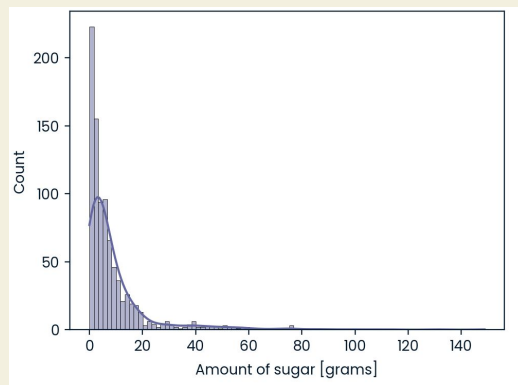
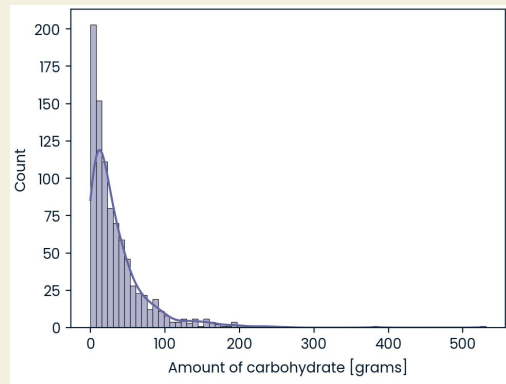
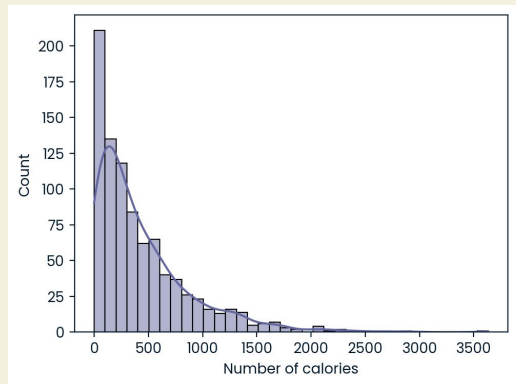
Food categories proportions



Probability of each category to be
well-known

- Target classes: balanced, might not need additional techniques
- **Category** feature: balanced between each categories, not need combine rare ones
- **Vegetable**, **Potato** and **Pork**: have highest chances to be popular

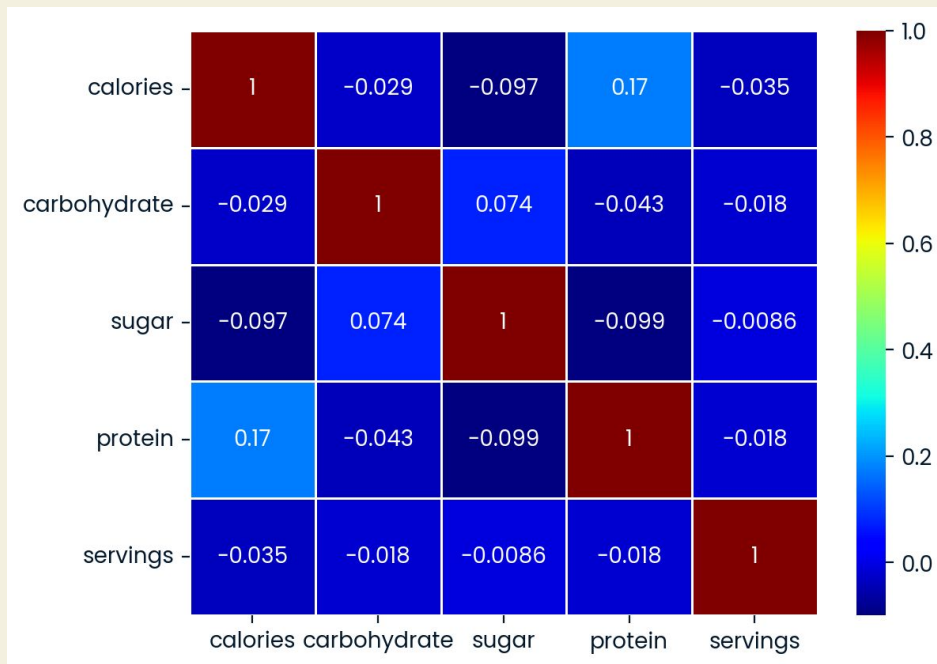
EDA [2/3]



Distributions of different numeric input features

- All features except **servings**: have long tail distributions
- Will need a suitable transformation before modeling

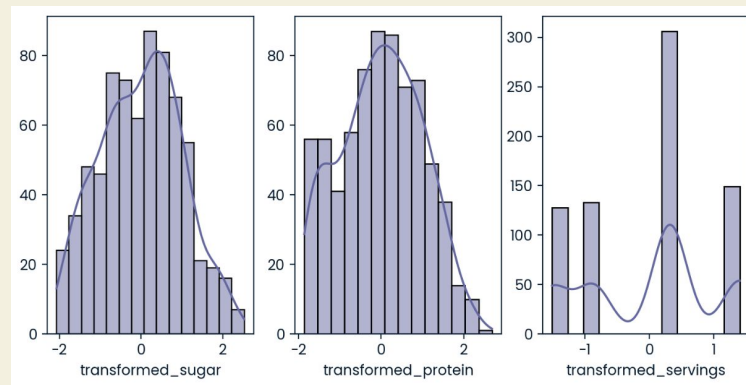
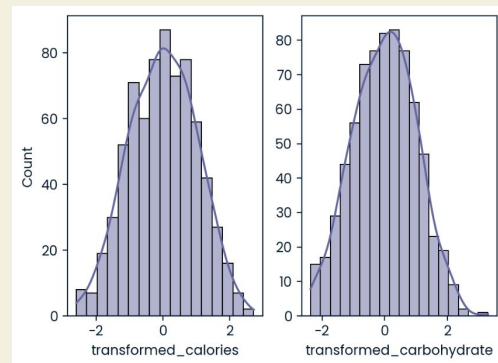
EDA [3/3]



- No strong correlations between numeric input features
- Keep all these features for modelling

Data Preprocessing

Column	Step 1	Step 2
calories	Power Transform	PCA
carbohydrate		
sugar		
protein		
servings		
category	One hot encoder	



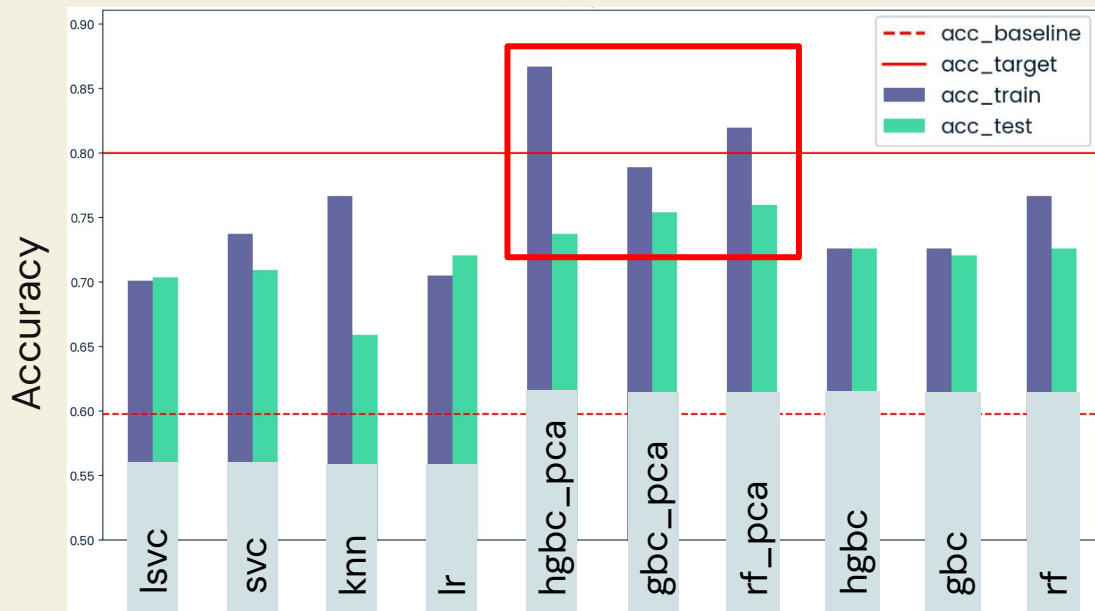
- Power transform \rightarrow Standardization
- One hot encoder \rightarrow categorical variables into a binary format
- PCA \rightarrow Dimension reduction

Models Construction

Model Type	Model	With PCA
Naive	Baseline	-
Linear	LinearSVC	Yes
	LogisticRegression	
Non-linear	SVC	
	KNeighborsClassifier	Yes and No
Ensemble	RandomForestClassifier	
	GradientBoostingClassifier	
	HistGradientBoostingClassifier	

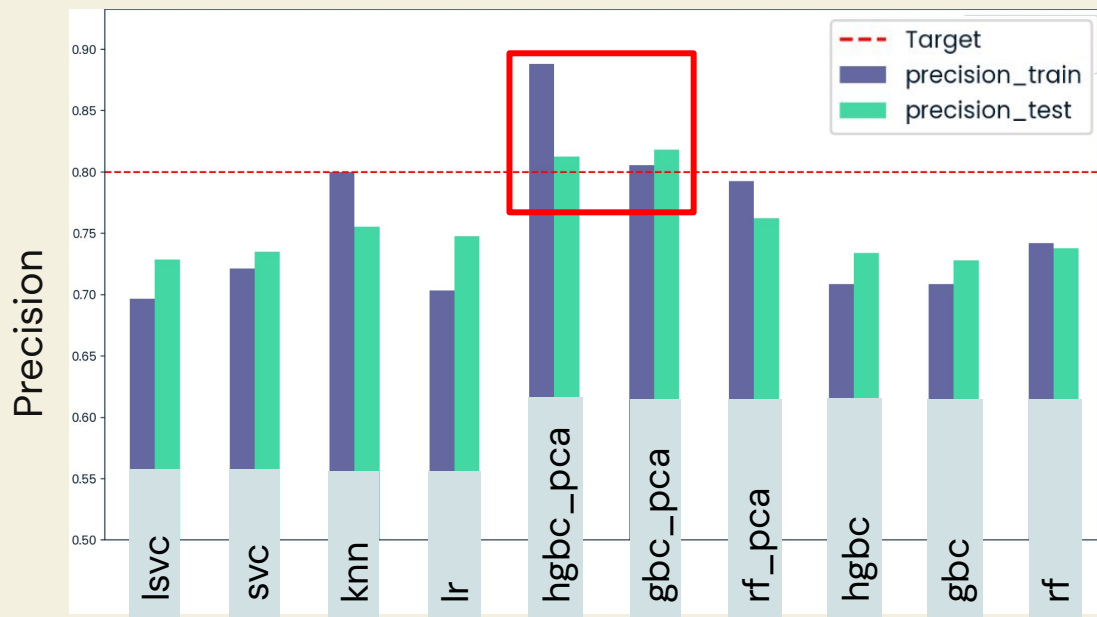
→ Apply Random Search for **Models' hyperparameters** and **Number of PCA's components**

Models Comparison



- All models: better than the baseline one
- Ensemble models w/ PCA: the best ones
- No models achieve 80% accuracy on the test dataset

Models Comparison



→ Gradient boosting and Histogram gradient boosting w/ PCA: achieve 80% accuracy on the number of positive predictions

4. Key Findings



4. Key Findings

- Top performers: `rf_pca`, `gbc_pca`, and `hgbc_pca` approach or exceed the target
- Overfitting: `rf_pca` and `hgbc_pca` show train-test gaps, suggesting overfitting.
- Underperformers: `knn`, `lsvc`, and `lr` perform just above baseline

→ PCA-enhanced ensemble models generalize best

→ `gbc_pca` stands out as the most balanced and robust across train and test sets.

5. Recommendations



5. Recommendations

✓ Deploy `gbc_pca` (i.e., gradient boosting classifier enhanced w/ PCA)

- Meets/exceeds 0.80 precision target
- Low overfitting (small train-test gap)
- Strong and stable test accuracy

🔧 Next Steps

- **Threshold Tuning:** Use precision-recall curve to reduce false positives
- **Post-Prediction Check:** Add business rules (e.g., ingredient trends, seasonality)
- **Monitoring:** Regular evaluation & retraining as user behavior shifts
- **PCA:** Keep in pipeline — it's boosting precision, but monitor over time