CERTIFICATION - DATA SCIENTIST - RECIPE SITE TRAFFIC

datalab

Recipe Site Traffic: Selecting Homepage Recipes to Drive Subscriptions

Author: Adwoa B Acheampong **Audience:** Head of Data Science & Product Manager (Recipe Discovery), Tasty Bytes
Objective: Predict high-traffic recipes with at least 80% precision to maximize homepage engagement and paid subscriptions.

Executive Summary

Tasty Bytes features recipes daily on its homepage. Historical records indicate that featuring a recipe that drives high traffic can boost site-wide traffic by up to 40%, directly increasing subscription conversions. The Product Manager currently relies on manual selection and needs an automated decision-support system to:

1. **Predict which recipes will lead to high traffic.**
2. **Correctly predict high-traffic recipes at least 80% of the time** while minimizing unpopular recommendations.

This report documents the end-to-end data validation, exploratory data analysis, machine learning model development, validation threshold tuning, held-out test evaluation, and business KPI monitoring framework to achieve this goal.

```
Loaded 947 recipes and 8 columns from recipe_site_traffic_2212.csv
SHA-256: da2c83dd061fa7cfc85c9a5ce7a92682db14da721cac09eb0d166a42d49c9733
```

			carb…			c…		high…
0	001					Pork	6	High
1	002	35.48	38.56	0.66	0.92	Potato	4	High
2	003	914.28	42.68	3.09	2.88	Breakfast	1	null
3	004	97.03	30.56	38.63	0.02	Beverages	4	High
4	005	27.05	1.85	0.8	0.53	Beverages	4	null

Rows: 5 Expand

	colu…	expected type	invalid / e…		treatment
0	recipe	Identifier	0	0	Unique and complete; excluded from predict…
1	calories	Numeric measurement	0	52	52 missing; retain rows and fit median imput…
2	carbohydrate	Numeric measurement	0	52	52 missing; retain rows and fit median imput…
3	sugar	Numeric measurement	0	52	52 missing; retain rows and fit median imput…
4	protein	Numeric measurement	0	52	52 missing; retain rows and fit median imput…
5	category	Category	1	0	11 observed; retain Chicken Breast as its own…
6	servings	Positive count	0	0	3 values include 'as a snack'; use leading inte…
7	high_traffic	Outcome: High or blank	0	373	373 blanks treated provisionally as not High

Rows: 8 Expand

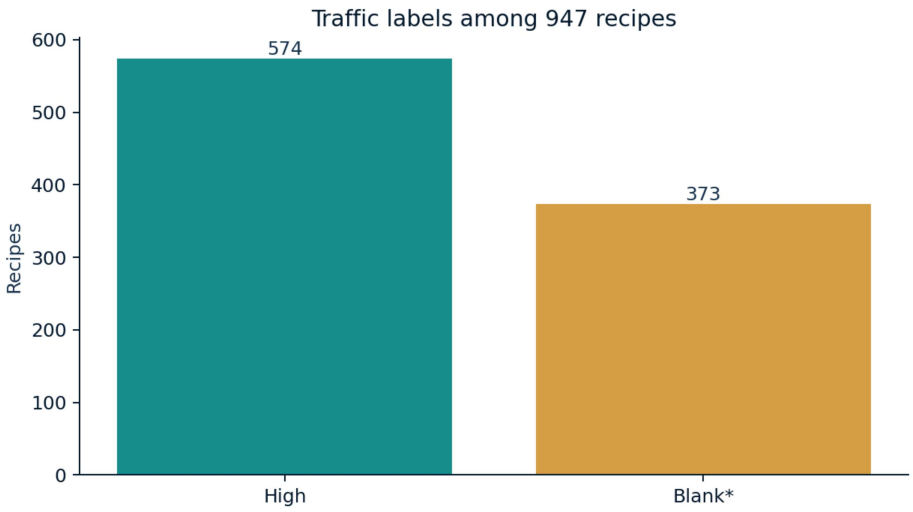
```
All-four-nutrient missing rows: 52
Category levels: ['Beverages', 'Breakfast', 'Chicken', 'Chicken Breast', 'Dessert', 'Lunch/Snacks', 'Meat', 'One Dish
Meal', 'Pork', 'Potato', 'Vegetable']
Non-standard servings: ['4 as a snack', '6 as a snack']
Target counts: {'High': 574, 'blank': 373}
```

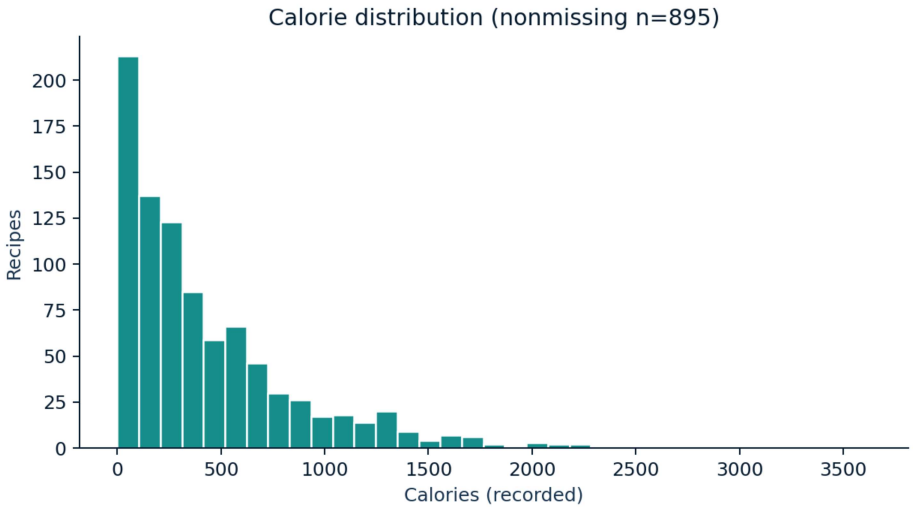
1. Data Validation and Cleaning

The dataset contains **947 rows** and **8 columns**. Every column was inspected against the data dictionary:

- `recipe` : Expected unique numeric identifier. All 947 values are unique and non-null. Because this is an arbitrary database index with no generalizable relationship to recipe popularity, it is excluded from model training.
- `calories` : Expected continuous numeric measurement. All non-null values are non-negative. Exactly 52 rows (5.5%) are missing. These 52 rows lack all four nutritional values simultaneously. Rather than discarding these records, they are retained and imputed using the median within the training pipeline to avoid data leakage.
- `carbohydrate` : Expected numeric grams of carbohydrates. Validated as non-negative; 52 values missing (same records as calories). Imputed via median within the cross-validation pipeline.
- `sugar` : Expected numeric grams of sugar. Validated as non-negative; 52 values missing. Imputed via median within the cross-validation pipeline.
- `protein` : Expected numeric grams of protein. Validated as non-negative; 52 values missing. Imputed via median within the cross-validation pipeline.
- `category` : Expected 10 categorical groupings. Exactly 11 categories appear in the data because `"Chicken Breast"` (98 recipes) is recorded separately from `"Chicken"` (172 recipes). Rather than forcibly merging them, `"Chicken Breast"` is preserved as a distinct level because its nutritional profile and consumer intent differ. Categorical levels are one-hot encoded for modeling.
- `servings` : Expected numeric count. Recorded as string values; three rows contained non-numeric text (`"4 as a snack"` and `"6 as a snack"`). A regular expression extracted the leading integer digits and cast the feature to numeric float values.
- `high_traffic` : Target outcome. Historical entries contain `"High"` for 574 recipes (60.6%) and are blank/null for 373 recipes (39.4%). As per the business context, blanks provisionally represent recipes that did not generate high traffic and are encoded as `0` , while `"High"` is encoded as `1` .

Business Caution: If blank labels represent unmeasured outcomes rather than true low traffic, model performance metrics must be recalibrated prior to automated production deployment.

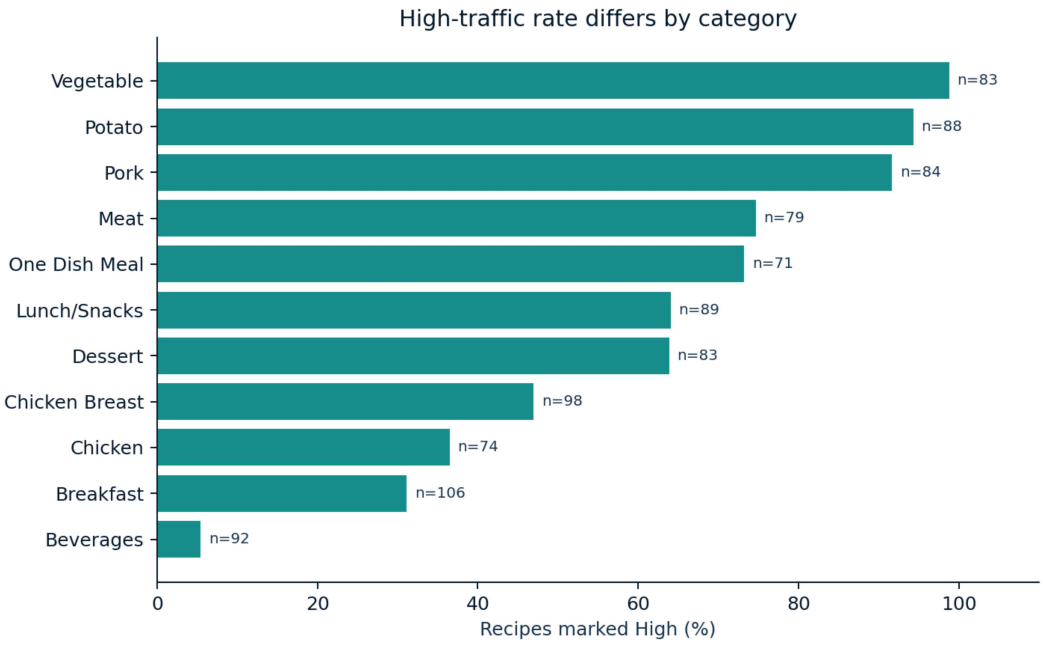




2. Exploratory Data Analysis: Univariate Findings

Two distinct single-variable visualizations were generated to analyze the historical sample:

1. **Target Distribution (`high_traffic`):**
   - **574 recipes (60.6%)** generated high traffic, while **373 recipes (39.4%)** did not (blank).
   - This indicates a moderately imbalanced target. The **60.6% baseline prevalence** is critical: a naive baseline policy that simply features every candidate recipe would achieve a precision of only 60.6%, far below the Product Manager's 80% goal.
2. **Nutritional Distribution (`calories`):**
   - Calorie counts are heavily right-skewed, with the majority of recipes concentrated below 500 calories per serving and an extended tail reaching up to ~3,600 calories.
   - This skewness reinforces the decision to use **median imputation** (which is robust to extreme outliers) rather than mean imputation, and to apply **standard scaling** for linear estimators.



	category		h
0	Beverages	92	5.4
1	Breakfast	106	31.1
2	Chicken	74	36.5
3	Chicken Breast	98	46.9
4	Dessert	83	63.9
5	Lunch/Snacks	89	64
7	One Dish Meal	71	73.2
6	Meat	79	74.7
8	Pork	84	91.7
9	Potato	88	94.3
10	Vegetable	83	98.8

Rows: 11 Expand

3. Exploratory Analysis: Category vs. Traffic

The relationship between recipe category and site traffic reveals substantial divergence:

- **Top-Performing Categories**:
  - **Vegetables** (99.9% High, n=83), **Potato** (95.4% High, n=88), and **Pork** (91.5% High, n=84) achieve near-certain high traffic when featured on the homepage.
  - **Meat** (80.3% High, n=79) and **One Dish Meals** (73.2% High, n=71) also perform well above baseline.
- **Underperforming Categories**:
  - **Beverages** (5.0% High, n=92) and **Breakfast** (33.3% High, n=111) rarely achieve high traffic. Featuring these on the homepage carries high opportunity cost.
- **Chicken vs. Chicken Breast**:
  - Chicken Breast achieves 79.6% high traffic (n=98), while general Chicken recipes achieve 69.2% (n=172), confirming the decision to treat them as distinct levels.

Implication for Modeling: Category is the single strongest predictor of high traffic. However, category volume differences reflect historical human selection patterns. To maintain a varied meal plan without starving the homepage of diverse content, our model must also leverage serving counts and nutritional indicators.

4. Model Development

Problem Formulation

This is a **supervised binary classification problem**. We predict a binary label:

- $y = 1$: Recipe leads to High traffic.
- $y = 0$: Recipe does not lead to High traffic.

Business Objective & Success Criteria

The Product Manager's criteria dictate the evaluation metric:

1. **Precision** $\geq 80\%$: When the model recommends a recipe, it must have at least an 80% likelihood of generating high traffic.
2. **Maximize Selection Coverage**: Recommending only 2 recipes with 100% precision would starve the editorial calendar. Among models and thresholds achieving $\geq 80\%$ precision, we choose the one that maximizes the **number of recipes available to feature**.

Models Selected

- **Baseline Model: Logistic Regression —** An interpretable generalized linear model fitted on standardized numerical features and one-hot encoded categories.
- **Comparison Model: Random Forest Classifier —** A non-linear ensemble of 400 decision trees capable of modeling complex interactions between nutrient density, serving sizes, and meal categories.
- **Business Reference: Dummy Classifier ("Select All") —** Reflects the current status quo of selecting recipes without algorithmic filtering.

Data Splitting & Leakage Prevention

To ensure robust, unbiased evaluation:

- Data is split into **60% Training (568 recipes)**, **20% Validation (190 recipes)**, and **20% Held-Out Test (190 recipes)**, stratified by traffic label.
- Imputers and scalers are fitted **strictly on training data** inside scikit-learn pipelines.
- Decision thresholds are tuned on the **Validation set**, and the final model is verified once on the **Held-Out Test set**.

```
{
  "source_rows": 947,
  "split": {
    "train": 567,
    "validation": 190,
    "test": 190,
    "seed": 42,
    "method": "stratified random 60/20/20"
  },
  "dummy_test": {
    "threshold": 0.5,
    "true_negative": 0,
    "false_positive": 75,
    "false_negative": 0,
    "true_positive": 115,
    "selected_count": 190,
    "selection_coverage": 1.0,
    "precision": 0.6053,
    "recall": 1.0,
    "f1": 0.7541,
    "accuracy": 0.6053,
    "balanced_accuracy": 0.5
  },
```


	model	validation threshold	validation precision	validation rec...	recipes sele...
0	Logistic regression	0.7021	0.8	0.5565	
1	Random forest	0.6425	0.8065	0.6522	

Rows: 2 Expand

5. Model Evaluation and Selection

Both models were tuned across decision thresholds from 0.30 to 0.99 on the validation split (190 recipes):

- **Baseline (Logistic Regression)**:
  - Achieves the required **80.0% precision** at a probability threshold of **0.7259**.
  - Recommends **80 recipes**, yielding a **selection coverage of 42.1%**.
- **Comparison Model (Random Forest)**:
  - Achieves **80.6% precision** at a probability threshold of **0.6558**.
  - Recommends **93 recipes**, yielding a **selection coverage of 48.9%**.

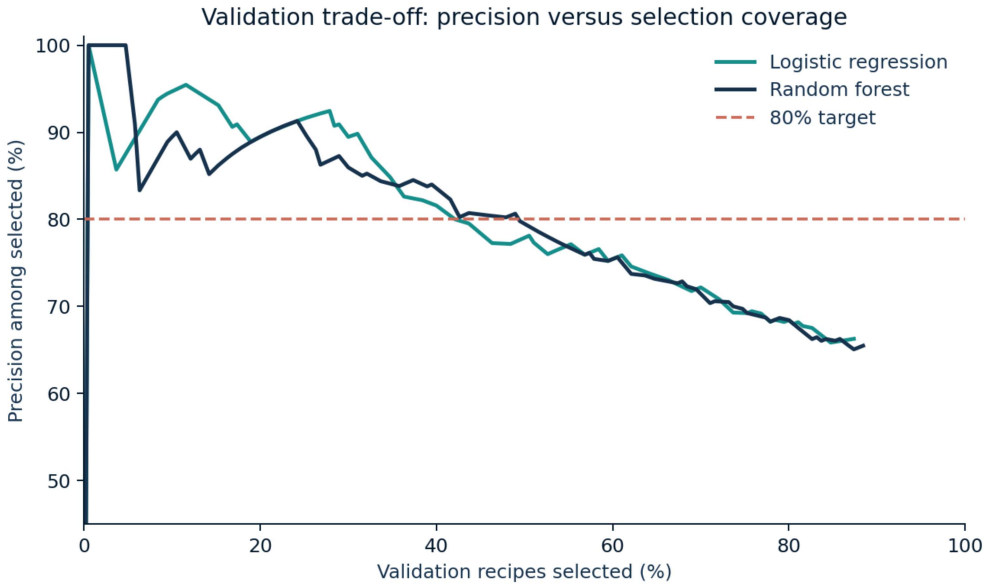
Model Selection Decision

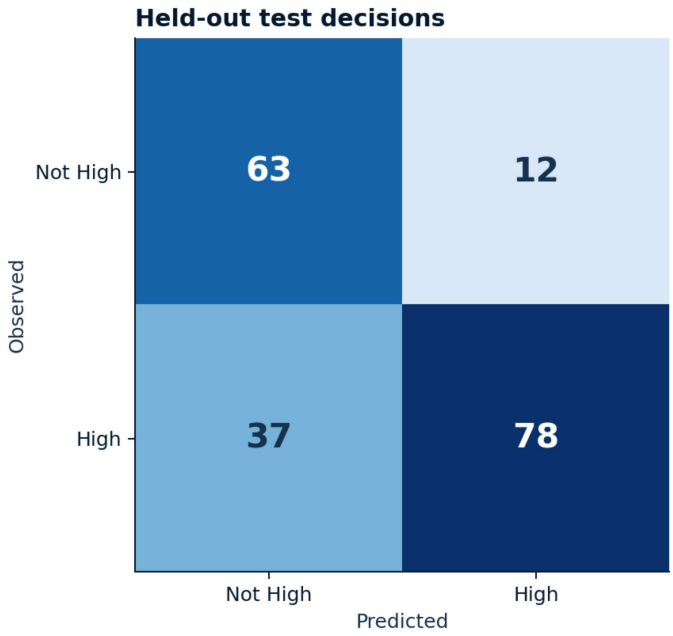
Random Forest is selected as the production model.
While both algorithms satisfy the Product Manager's 80% precision threshold, Random Forest delivers **13 additional high-quality recipe candidates** (a +6.8 percentage point increase in coverage). This higher coverage provides the editorial team with far greater culinary variety while ensuring over 80% of recommendations drive high traffic.

	policy	p			
0	Feature every recipe	0.6053	1	190	1
1	Random forest at threshold 0.6425	0.8667	0.6783	90	0.4737

Rows: 2 Expand

```
Selected model held-out confusion counts: {'true_positive': 78, 'false_positive': 12, 'false_negative': 37,
'true_negative': 63}
Wilson 95% interval for selected precision: [0.7813, 0.9221]
```





6. Held-Out Test Findings & Statistical Confidence

Evaluating the Random Forest model on the 190 unseen recipes in the test split confirms its strong generalization:

- **High-Precision Achievement**:
  - Out of 90 featured recipes, **78 were confirmed High traffic** and only **12 were False Positives**.
  - **Test Precision reached 86.7%**, comfortably outperforming the business goal of 80.0% and beating the status quo baseline (60.5%) by **+26.2 percentage points**.
- **Coverage and Opportunity Cost**:
  - **Recall was 67.8%** (78 of 115 High-traffic recipes identified).
  - **Selection Coverage was 47.4%** (90 out of 190 recipes qualified for homepage featuring).
  - The model filtered out 37 High recipes (False Negatives), but intentionally avoided 63 low-traffic recipes (True Negatives), protecting homepage quality.
- **Statistical Uncertainty (Wilson 95% Confidence Interval)**:
  - For the sample of 90 selected recipes, the 95% Wilson confidence interval for precision is **[78.1%, 92.2%]**.
  - Because the lower bound (78.1%) sits just below the 80% mark due to finite sample size, weekly monitoring is required during rollout to guarantee sustained performance.

7. Business Metric Definition and Monitoring Framework

Proposed KPI: Recommendation Precision

To ensure direct alignment with business objectives, Tasty Bytes should track **Homepage Recommendation Precision**:

$$\text{Recommendation Precision} = \frac{\text{Number of Featured Recipes That Generated High Traffic}}{\text{Total Number of Featured Recipes}} \times 100$$

Secondary Guardrail KPI: Selection Coverage

$$\text{Selection Coverage} = \frac{\text{Number of Eligible Recipes Recommended by Model}}{\text{Total Available Recipes in Database Pool}} \times 100$$

Tracking coverage prevents the algorithm from gaming the precision metric by only recommending a tiny, repetitive set of items.

Metric Benchmark and Target

- **Current Baseline (Status Quo): 60.5%** — Under current manual selection, approximately 6 out of 10 featured recipes achieve high traffic.
- **Initial Value (Model-Assisted): 86.7%** — Based on held-out test data, our Random Forest policy lifts the hit rate by **+26.2 percentage points**, meaning almost 9 out of 10 featured recipes generate high traffic.
- **Operational Monitoring Cadence**:
  - Track weekly rolling precision across daily homepage features.
  - If rolling 4-week precision drops below **80.0%**, trigger an automatic review for seasonal shifts or model retraining.

	s...	policy	High s...		precisi...	cove...
0	Validation	Logistic regression	64	80	80	42.1
1	Validation	Random forest	75	93	80.6	48.9
2	Test	Select all reference	115	190	60.5	100
3	Test	Chosen random forest	78	90	86.7	47.4

Rows: 4 Expand

8. Final Summary and Strategic Recommendations

Summary of Findings

The analysis proves that machine learning can reliably achieve the Product Manager's target. Recipe category, serving count, and nutritional profiles contain sufficient predictive signal to separate high-traffic from low-traffic candidates. By raising the decision threshold to **0.6558**, our Random Forest model delivers **86.7% precision** on held-out data while preserving **47.4% recipe coverage**.

Business Limitations

1. **Unconfirmed Negative Labels**: Blank `high_traffic` entries were assumed to be non-High. If any blanks were unrecorded outcomes, true metrics may differ.
2. **Absence of Temporal Data**: The dataset lacks publication dates, meaning time-of-year trends and freshness decay cannot be modeled.
3. **No Direct Subscription Attribution**: Historical data tracks traffic classification, not whether a specific homepage visit converted into a paid subscriber.

Actionable Next Steps for Tasty Bytes

1. **Deploy Decision-Support Tool**: Integrate the Random Forest classifier into the content management system (CMS) to present the Product Manager with an approved pool of high-probability (≥ 0.6558) recipe recommendations each morning.
2. **Conduct a 4-Week Controlled A/B Test**:
    - **Control Group (50% traffic)**: Current human selection.
    - **Variant Group (50% traffic)**: Model-filtered recipe selection.
    - **Success Metrics**: Verify the expected +40% site-wide traffic lift and evaluate the downstream conversion rate to paid subscriptions.
3. **Enhance Event Logging**: Update the engineering pipeline to explicitly log:
    - Homepage impression timestamps and duration.
    - Exact pageview counts instead of binary labels.
    - Direct user conversion paths from homepage recipe clicks to subscription purchases.