Predicting Popular Recipes for Tasty Bytes

DataCamp Certification – Final Project

Yanze Liu

Project Context

The product team needs to predict which recipes on the homepage will generate high traffic.

Project Goal

Build a predictive model and deliver actionable business recommendations.

Our Mission: Pinpoint High-Traffic Recipes to Drive Engagement

The Product Manager's Request

- Predict which recipes will drive **"High Traffic"**.
- Achieve an **80% success rate** in identifying these popular recipes.
- Ultimately, **increase overall site traffic** and **subscription conversions**.

Our Plan of Action

- Validate and clean the **raw recipe data**.
- Explore the data visually to **uncover patterns**.
- Train and rigorously compare **predictive models**.
- Propose key metrics for **ongoing monitoring**.
- Deliver clear, **actionable recommendations**.

Forging a Reliable Dataset from Raw Ingredients

Numerical Features `calories`, `carbohydrate`, `sugar`, `protein`

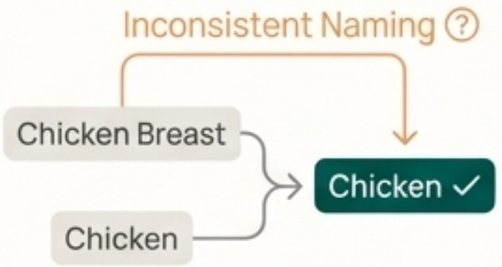
Challenge:: Small number of missing values (52 instances).

Solution: Imputed using the median to avoid skew from outliers.

Category Feature `category`

Challenge:: Inconsistent naming (e.g., 'Chicken Breast' vs. 'Chicken').

Solution:: Standardized names before one-hot encoding.



Servings Feature `servings`

Challenge:: Contained mixed text and numbers (e.g., '4 as a snack').

Solution:: Used regular expressions to extract only the numerical digits.

Target Variable `high_traffic`

Challenge:: Missing values (373 instances).

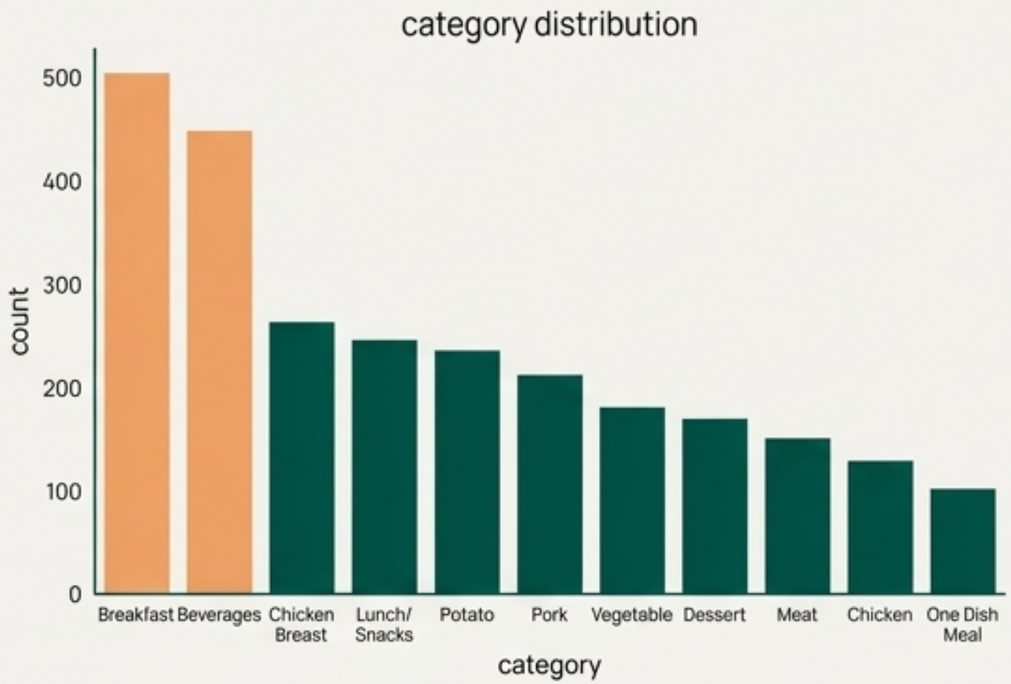
Solution:: Filled missing values as 'Low' and converted to a binary (1/0) format for modeling.

Initial Exploration Reveals Key Data Characteristics

Nutritional Features are Heavily Skewed

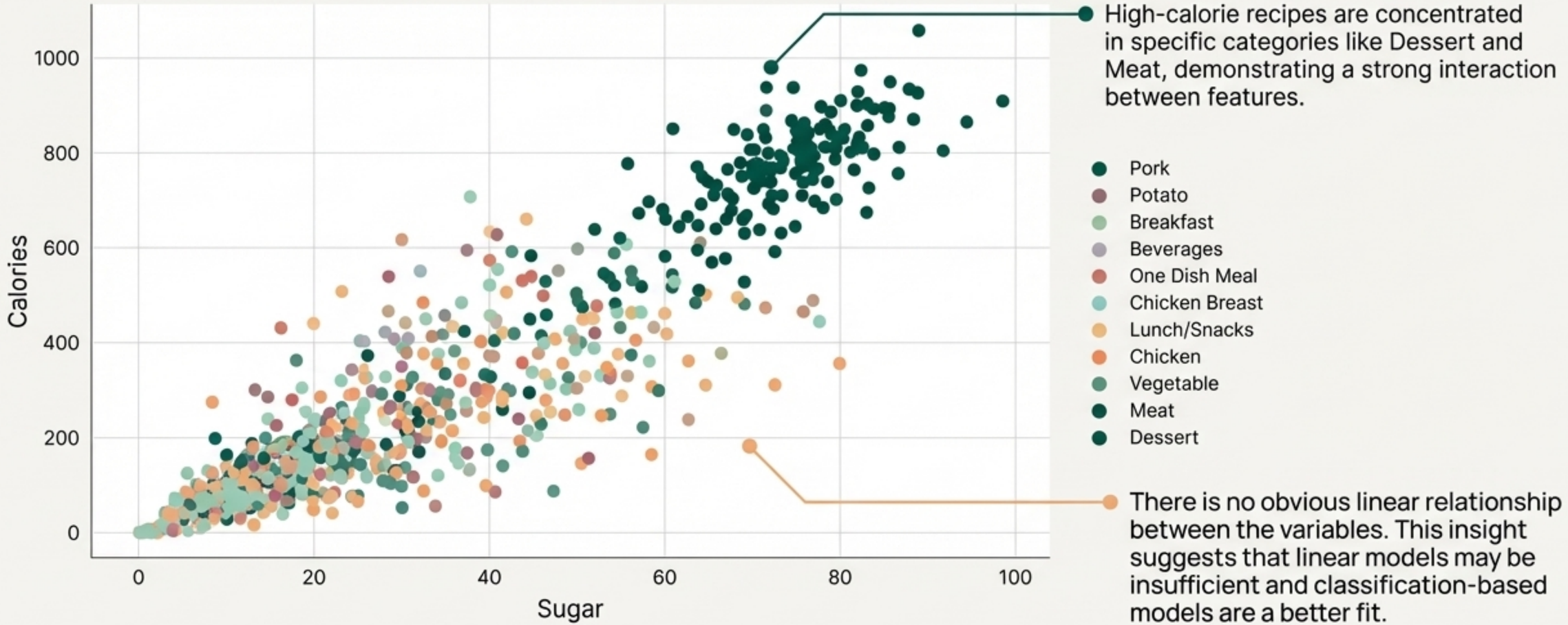
The vast majority of recipes have low calorie and sugar content, but a small number of recipes have extremely high values. This right-skew is consistent across all nutritional features.

Recipe Categories are Unbalanced



The dataset is not evenly distributed across categories. Breakfast and Beverage recipes are significantly more common than others.

Nutritional Profiles Differ Dramatically Across Recipe Categories



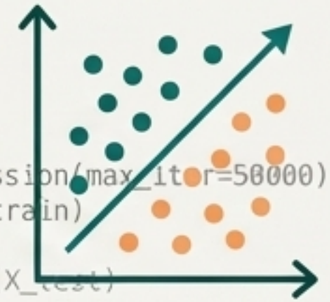
High-calorie recipes are concentrated in specific categories like Dessert and Meat, demonstrating a strong interaction between features.

There is no obvious linear relationship between the variables. This insight suggests that linear models may be insufficient and classification-based models are a better fit.

A Showdown: Pitting Interpretability Against Complexity

What's the best approach for this **Binary Classification** problem (predicting High vs. Low traffic)?

The Baseline

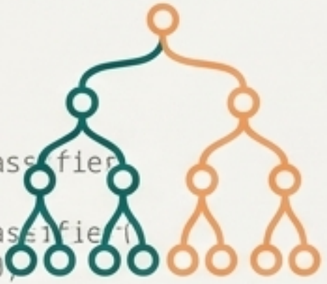


```
logit = LogisticRegression(max_iter=50000)
logit.fit(X_train, y_train)

y_red = logit.predict(X_test)

print("Accuracy:", accuracy_score( y_pred)
print("\nClassifi                    t(y_test, y_pred))
```

Logistic Regression

A robust and highly interpretable model that provides a strong, linear baseline for performance.

The Challenger



```
rf = RandomForestClassifier

rf = RandomForestClassifier(
    n_estimators=300,
    random_state=42,
    class_weight='balanced',
)

rf.fit(X_train, y_train)
```

Random Forest

A powerful, non-linear model capable of capturing intricate patterns and interactions between recipe features.

A Clear Winner Emerges: Simplicity and Stability Outperform Complexity

Metric	Logistic Regression	Random Forest
Accuracy	**~77%**	~73%
Recall (for 'High')	**~79%**	~76%
F1-Score	**~78%**	~73%

This table's data is derived from the classification reports.

The Logistic Regression model is the clear winner, delivering higher performance across all key metrics and bringing us closer to the business goal of 80%.

The Threshold Lever

By lowering the prediction threshold to 0.43, we can increase **Recall** to find *more* high-traffic recipes. This provides a strategic lever to balance capturing hits versus showing fewer low-traffic recipes.

The One Metric That Matters: Maximizing High-Traffic Recall

The Business Imperative

The greatest risk is failing to promote a recipe that could have been a hit.

High Recall minimizes this lost opportunity by ensuring we capture the maximum number of potential winners for the homepage.

RECALL

What percentage of *actual* high-traffic recipes did our model successfully identify?

Our Model's Performance

Our Logistic Regression model achieves a Recall of **79%**, right on on the cusp of the 80% target.

With threshold tuning, we can confidently **exceed this goal**.

A Strategic Roadmap to Higher Homepage Engagement

Our analysis confirms that we can reliably predict high-traffic recipes. The Logistic Regression model provides a strong foundation that is ready for real-world testing and future enhancement.

Actionable Next Step

Launch an A/B test. Use the model's predictions to populate a variant of the homepage and measure the direct impact on traffic and user engagement against the current version.

Future Improvements

Collect more powerful features like **preparation time, ingredient cost, number of ingredients, and user ratings.**

Experiment with Gradient Boosting models (e.g., XGBoost) for potential performance lifts.

Ongoing Success

Establish a continuous monitoring dashboard focused on High-Traffic Recall. Use the prediction threshold as a strategic tool to adapt to changing business priorities.