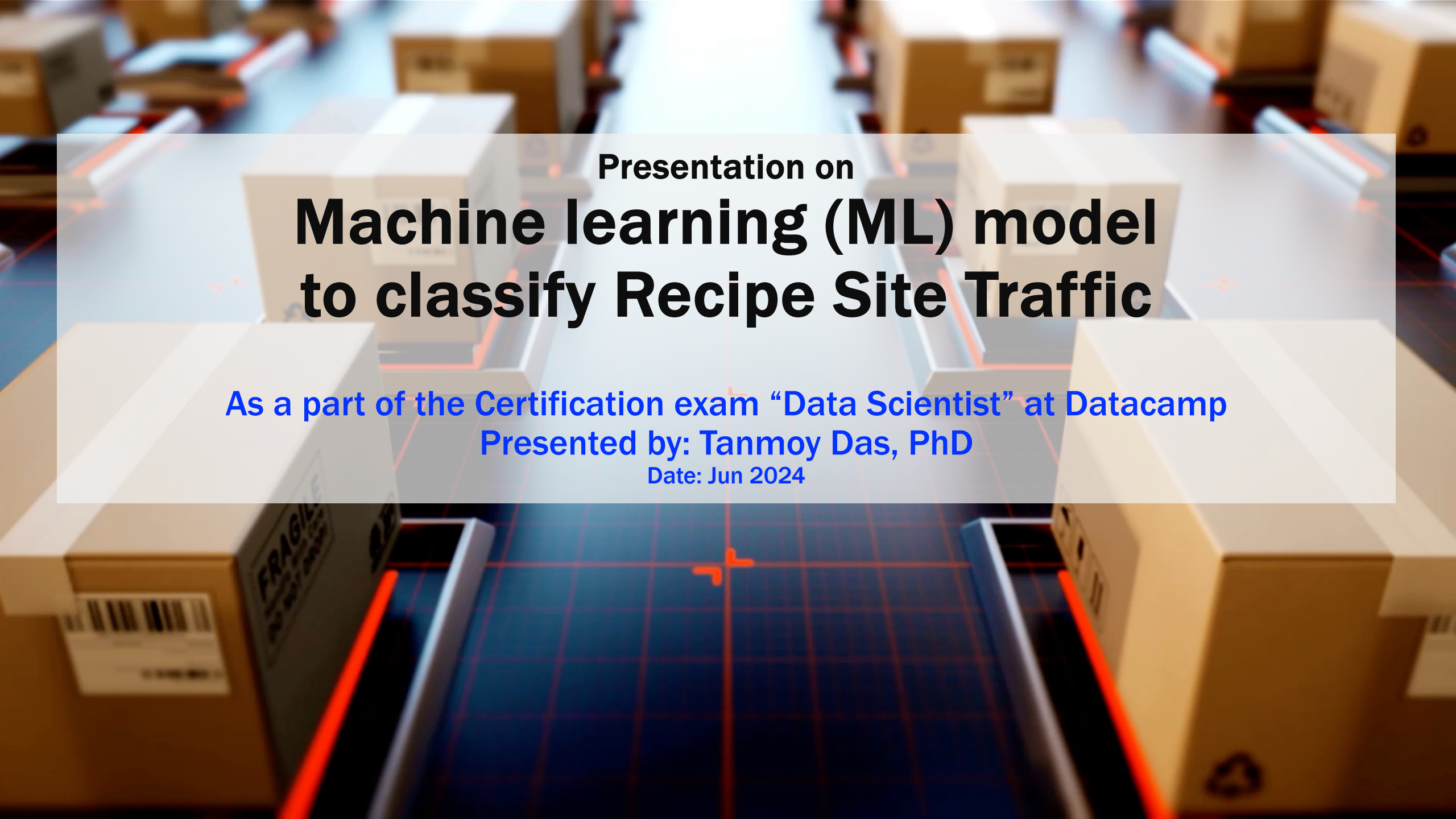

Presentation on

Machine learning (ML) model to classify Recipe Site Traffic

As a part of the Certification exam “Data Scientist” at Datacamp

Presented by: Tanmoy Das, PhD

Date: Jun 2024

Outline

1. Data Validation
2. Exploratory Analysis
3. Model Development
4. Model Evaluation
5. Business Metrics
6. Final summary statement

Data Validation & Exploratory Data Analysis (EDA)

- Validation
  - Datatype, shape
  - null values
  - Consistency in categorical variables
- Data cleaning
  - Feature and target variables are cleaned as necessary. For instance, category column contains 'Chicken' & 'Chiken Breast', both should belong to same category 'Chicken'.
  - Rows with null values are dropped from the dataset

Table1. Sample dataset

recipe	calories	carbohydrate	sugar	protein	category	servings	High traffic
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

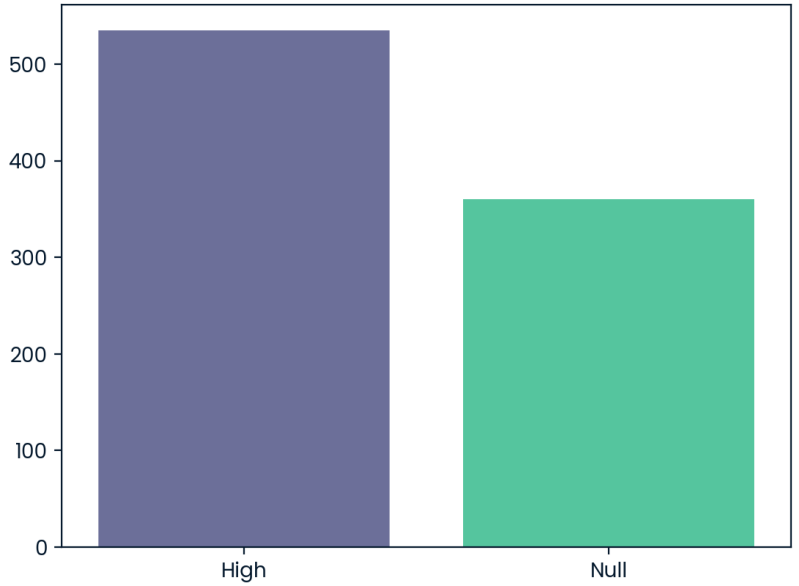

Fig1. Unique values in target variable

EDA continued...

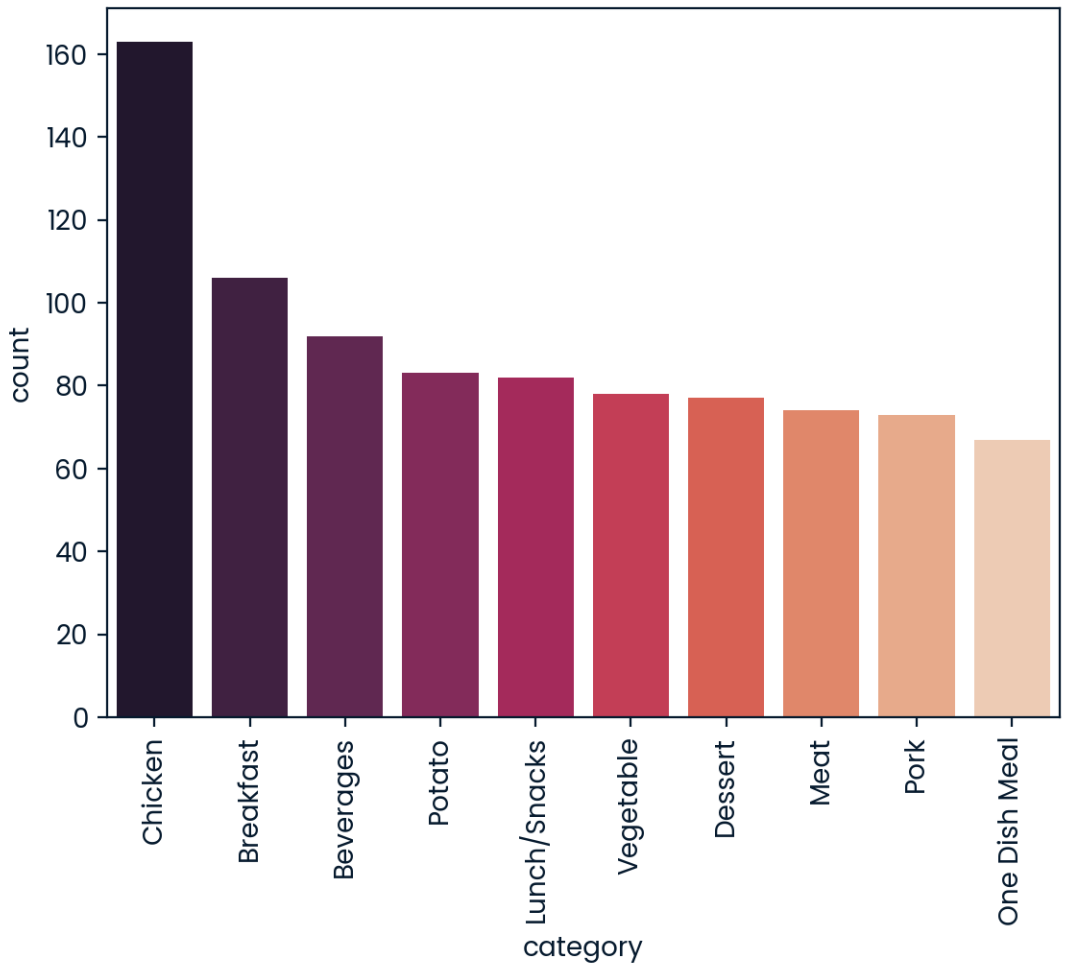

Fig2. Frequency of recipe category

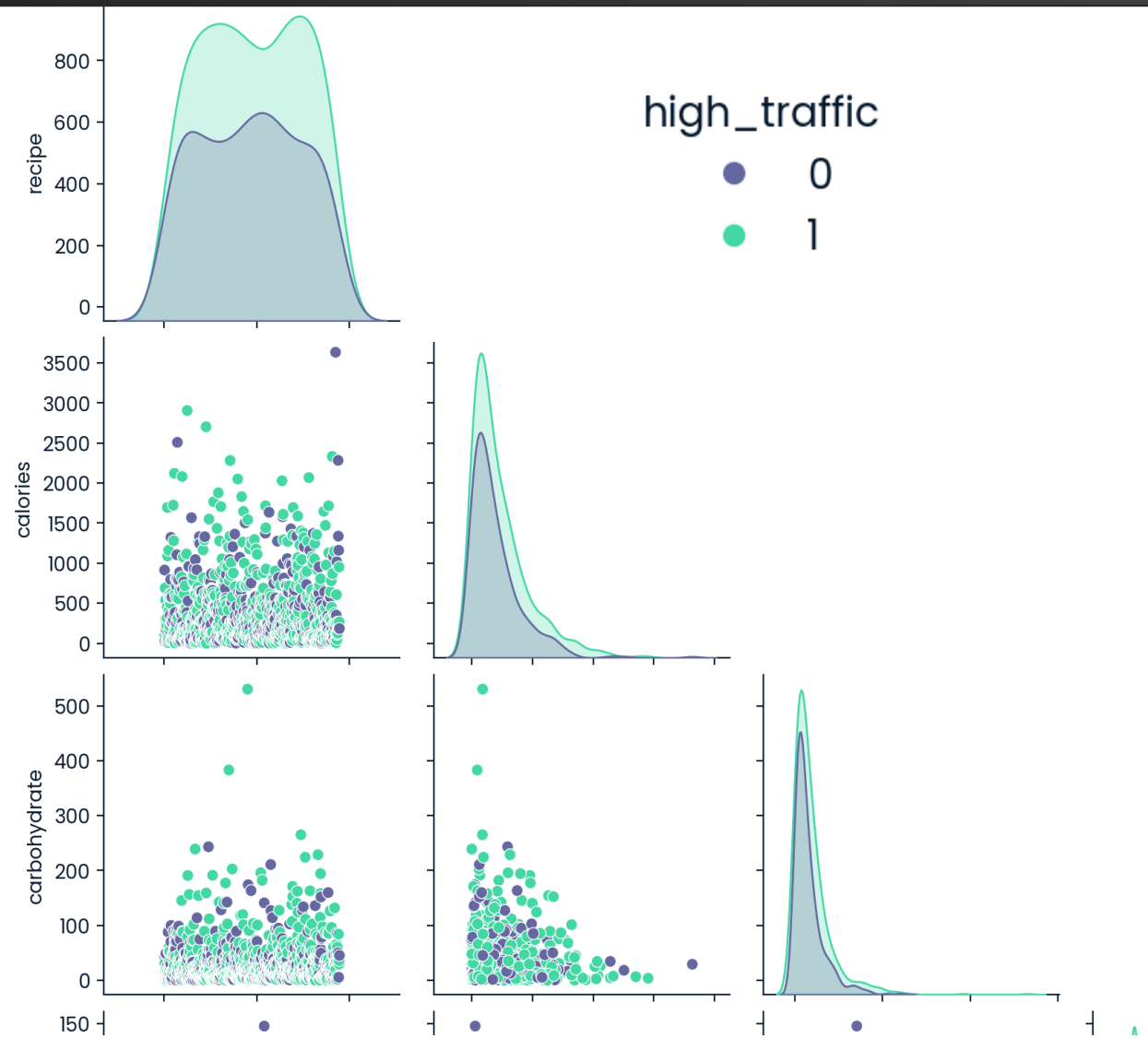

Fig3. Pair-plot of feature variables

Model Development

- **Logistic Regression** (base model)
  - Map the probability of a binary outcome (0 or 1) as a function of predictor variables.
  - The relationship is expressed using the logistic (sigmoid) function, which outputs probabilities between 0 and 1.
  - $z = \beta_0 + \beta_1 x_1 + \cdots \beta_n x_n$ where β is coefficient, and x is features
  - Logistic function, $f = \frac{1}{1+e^{-z}}$

- **Random Forest** (2nd model)
  - **Bootstrap Sampling**
  - **Tree Construction**
    - **Feature Selection**
    - **Node Splitting**
  - **Tree Growing**
  - **Aggregation**: Once all trees are constructed, combine their predictions:
    - For classification tasks, use majority voting where each tree casts a vote for the predicted class, and the class with the most votes is chosen.

Model Evaluation

- **Accuracy**

$$\frac{\text{Number of correct prediction}}{\text{Total number of prediction}}$$

- **Confusion Matrix**
  - For a binary classification, confusion matrix is a 2x2 table

	Actual Positive	Actual Negative
Predicted Positive	TP	FP
Predicted Negative	FN	TN

Table2. Model performance on test dataset

Model	Accuracy	Confusion Matrix
LR	0.75	$\begin{bmatrix} 51 & 22 \\ 19 & 87 \end{bmatrix}$
RF	0.69	$\begin{bmatrix} 42 & 31 \\ 23 & 83 \end{bmatrix}$

Business Metrics

- Classifying a high-traffic event as a low traffic can be damaging for our business. Hence we should be careful about false positives
- End-to-end development of a ML project is demonstrated
- Since category is most significant variable, lets take more caution while collecting data regarding this

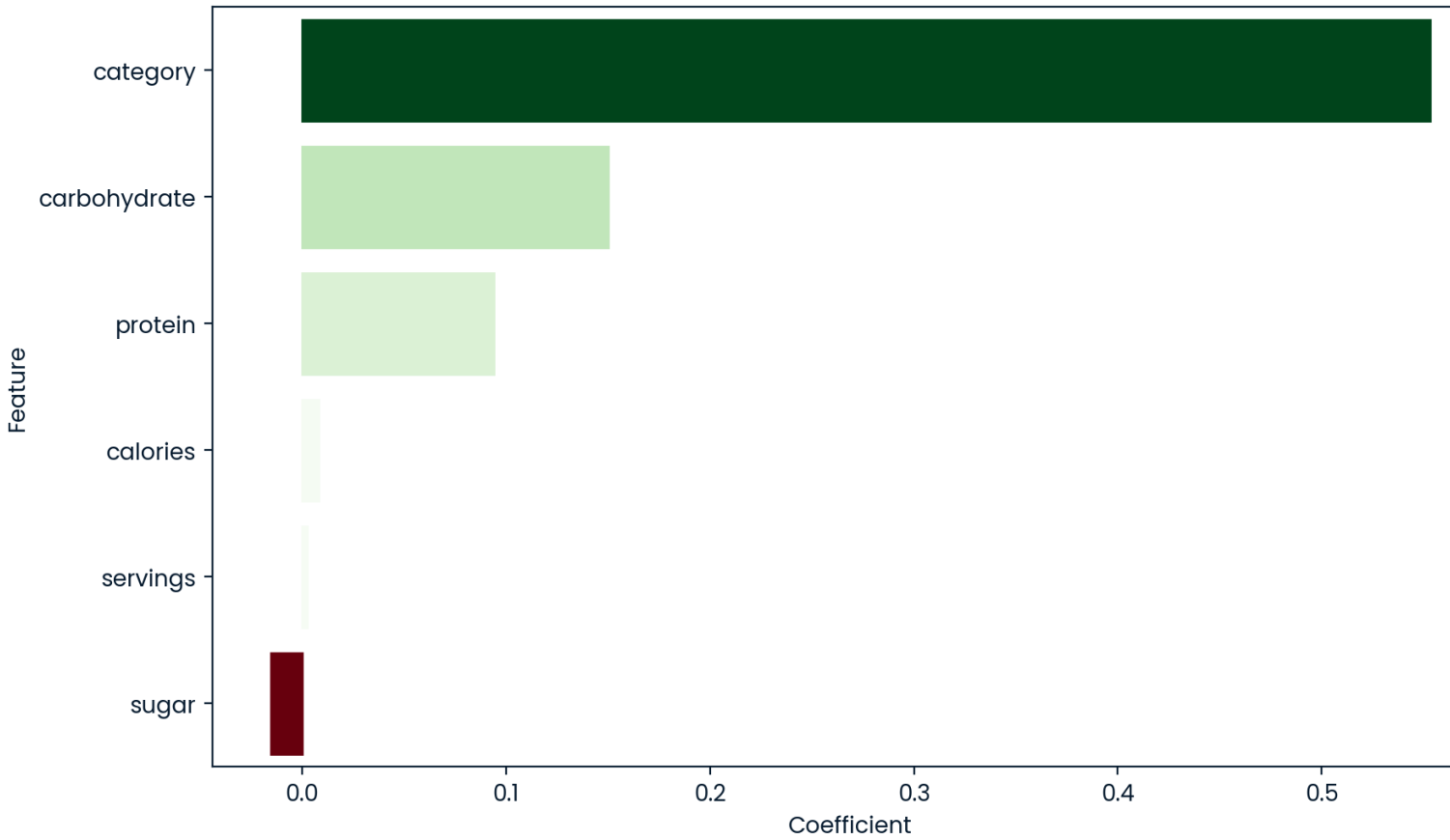

Fig4. Variable importance in Logistic Regression

Final Remark

- Logistic Regression performs better for our given test dataset (75% accurate compared to 69% by RF).
- Hence, logistic regression can be deployed in Azure for future recommendation of recipe with high traffic.