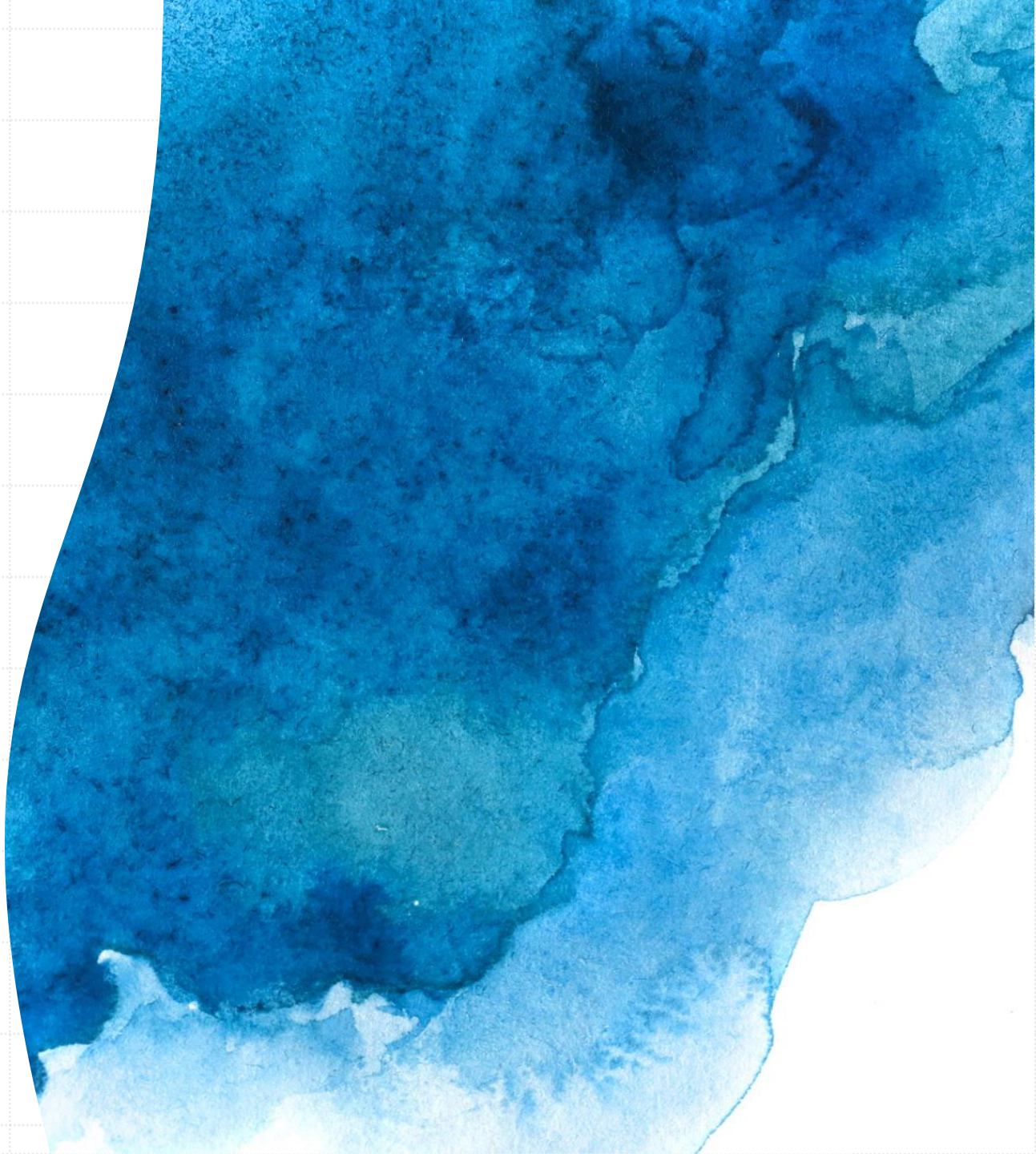
Will this Recipe be Popular?

Prepared for the Product team



Objectives

- ***Should this recipe be listed?***
- ***What kind of recipes have higher chance to become popular?***

Background Overview

- Popular Recipe drives higher (as much as 40%) traffic to the websites
- More traffic indicates more subscriptions
- Subscription is essential to our business
- Predict which recipes will lead to high traffic and correctly predict high traffic recipe 80% of the time

Our Solution

- Develop a predictive model to classify recipes as high or low traffic based on key features, and achieve at least 80% accuracy in correctly identifying high-traffic recipes.
- Identify key factors that drive high traffic to help the product team optimize recipe selection.
- Deploy the model for prediction, enabling the product team to assess a recipe's potential popularity before listing.

Dataset Overview and Data Preparation

Dataset Overview:

Our dataset consists of 947 observations, including the following features:

a) Recipe ID

b) Nutritional facts (calories, carbohydrates, sugar, and protein)

c) Recipe category

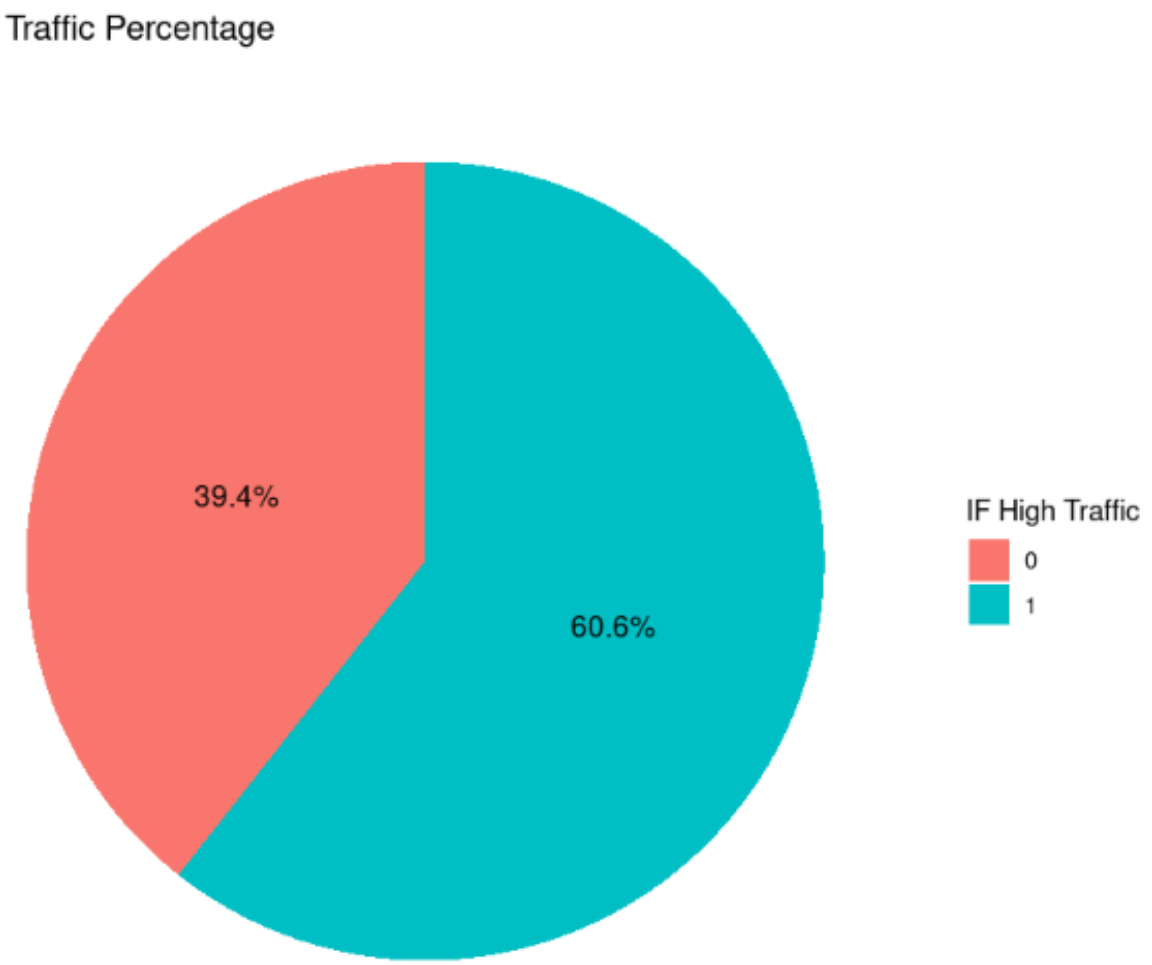
d) Number of servings per recipe

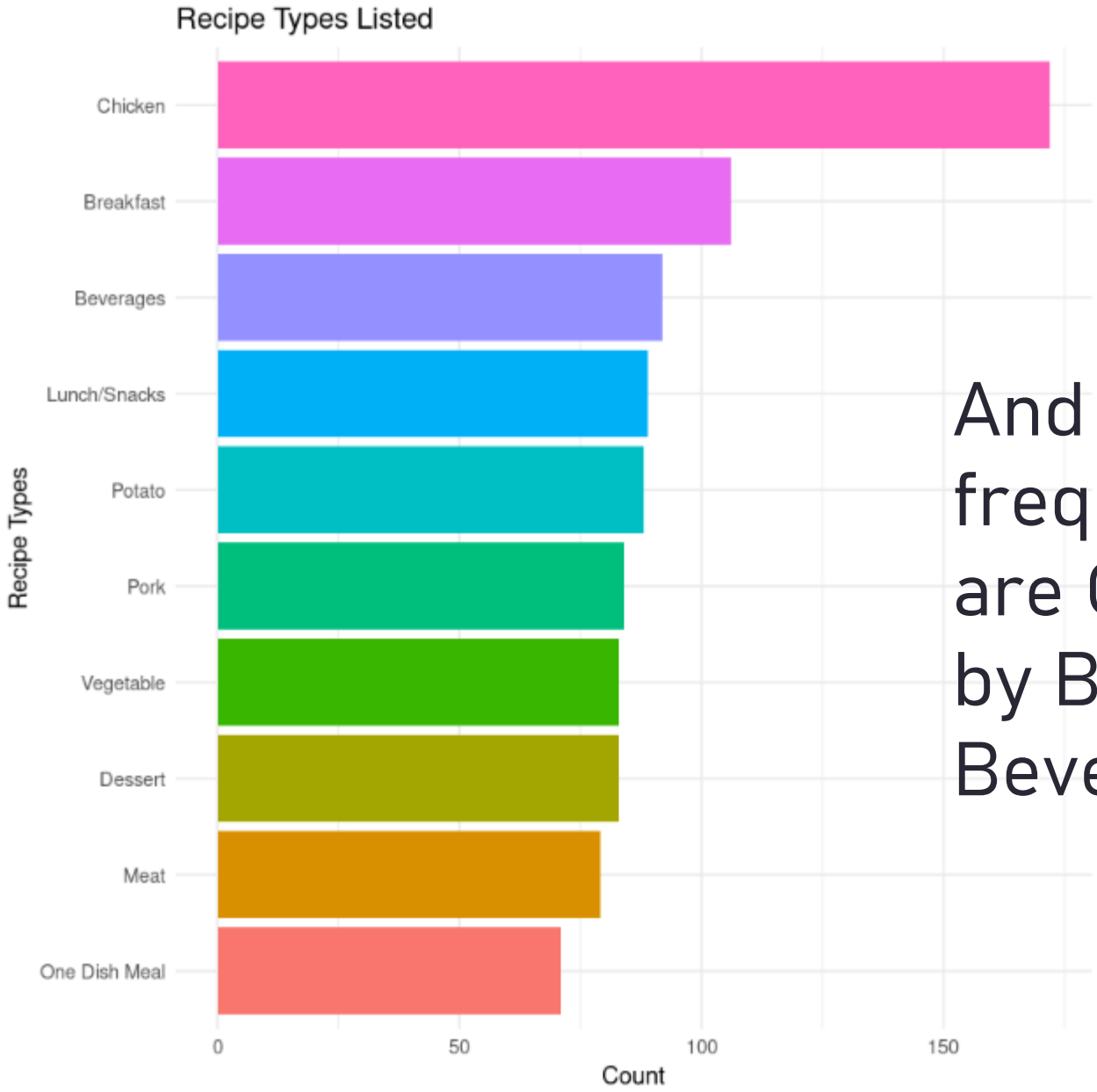
e) Traffic indicator (whether the recipe received high traffic)

Dataset Preparation:

1) Converted “servings” into numeric values (“4 as snack” to 4, “6 as snack” to 6)
2) Filled missing values in column calories, carbohydrates, sugar, and protein with their median
3) Merged “Chicken Breast” and “Chicken” into a single category: “Chicken”
4) Transformed “high_traffic” into an integer variable: 1 for “High” traffic, 0 for missing values (low traffic)
5) Training and testing prop. is 80% : 20%, stratified sampling on high_traffic is used

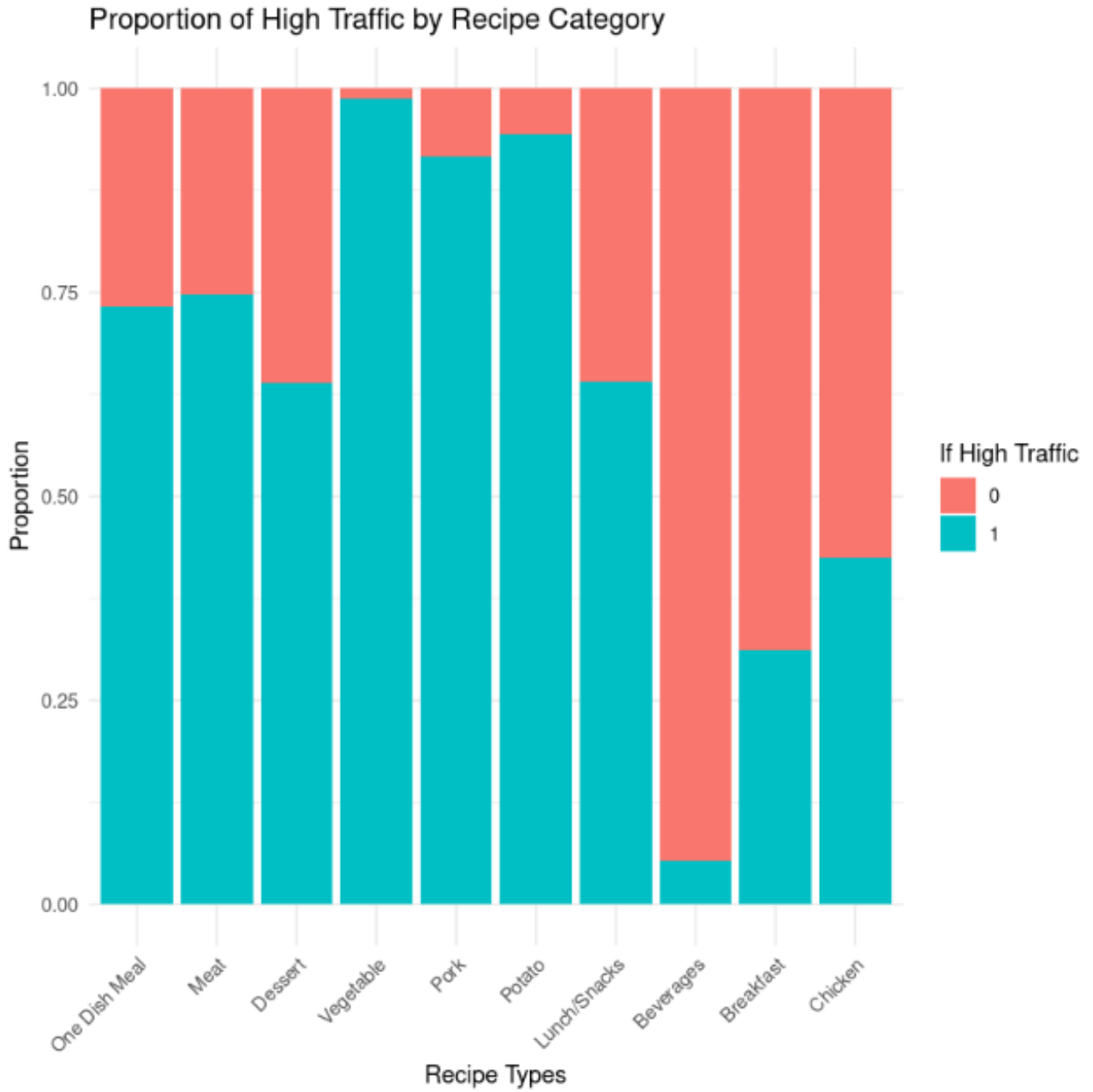
In this dataset, 60.6% of the recipes received high traffic

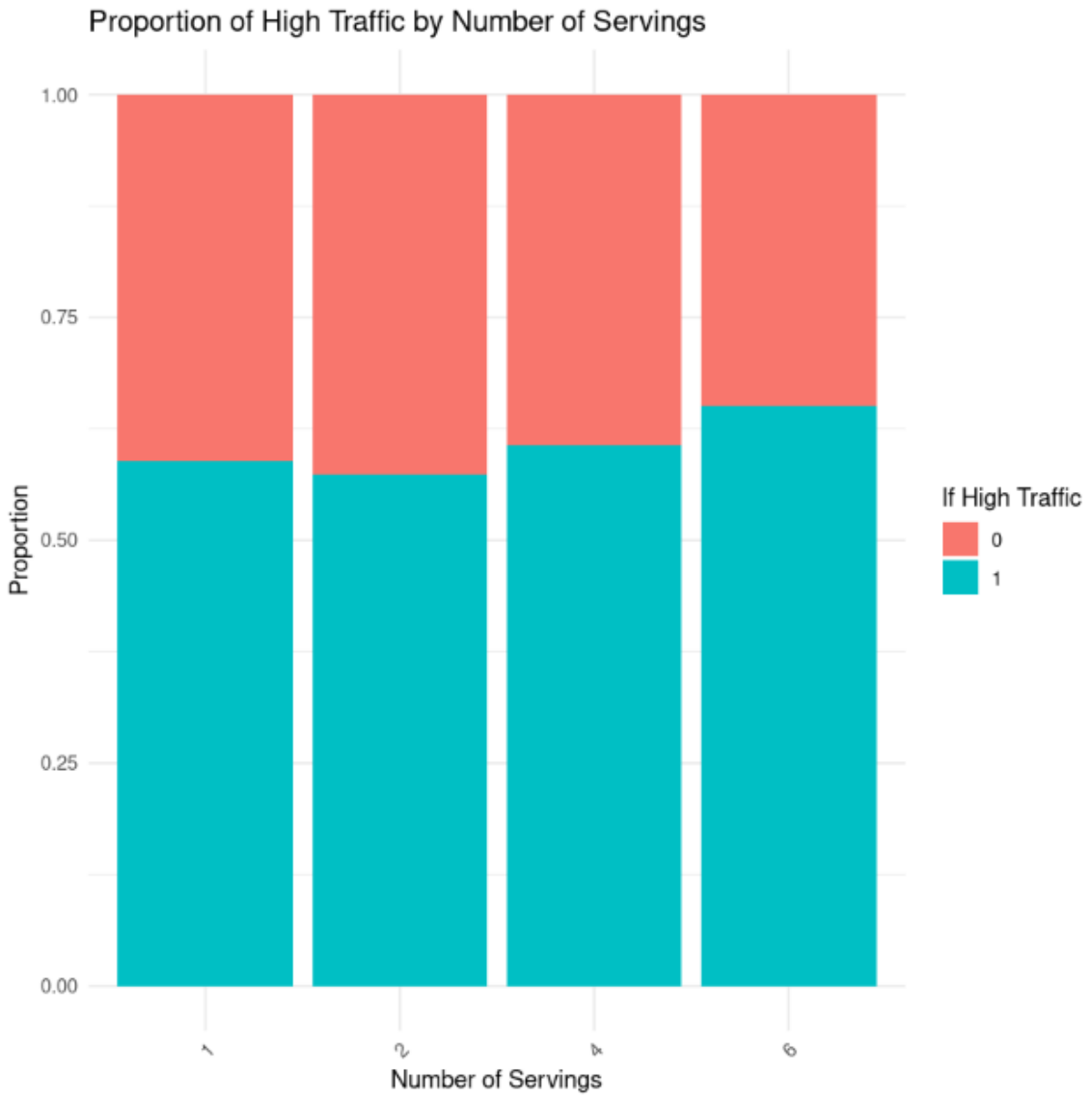




And The most frequently listed recipes are Chicken, followed by Breakfast and Beverage

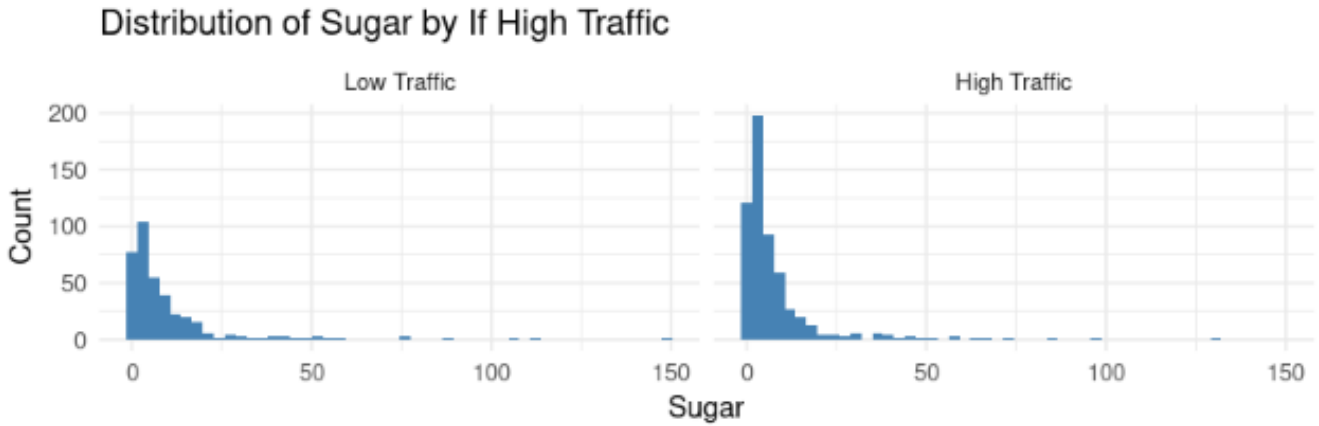
However, Vegetable, Pork, and Potato recipes have a higher proportion of high-traffic instances, Beverage recipes have the lowest likelihood of gaining high traffic

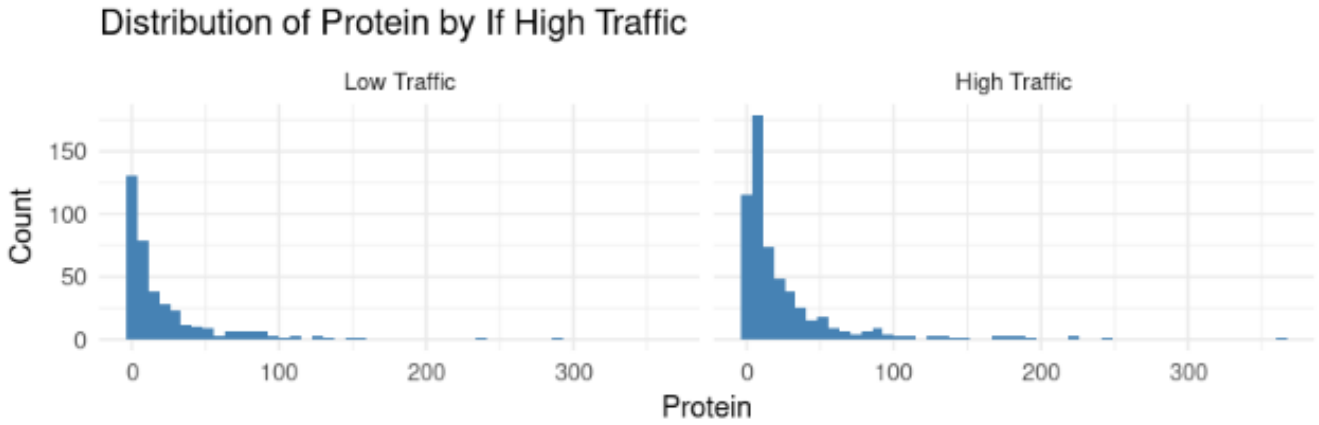




The recipes that serve 6 people have a marginally higher chance of becoming popular

The distributions of calories, sugar and protein differ only slightly between high and low traffic recipes

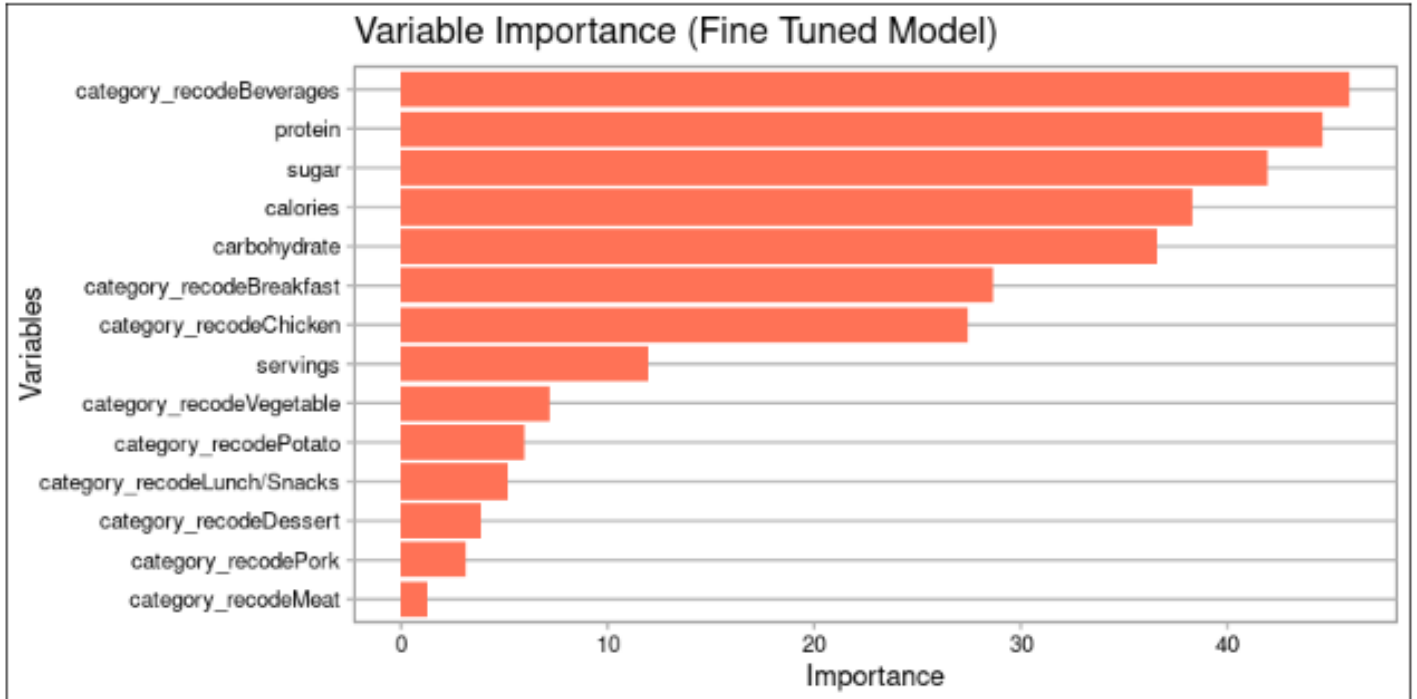


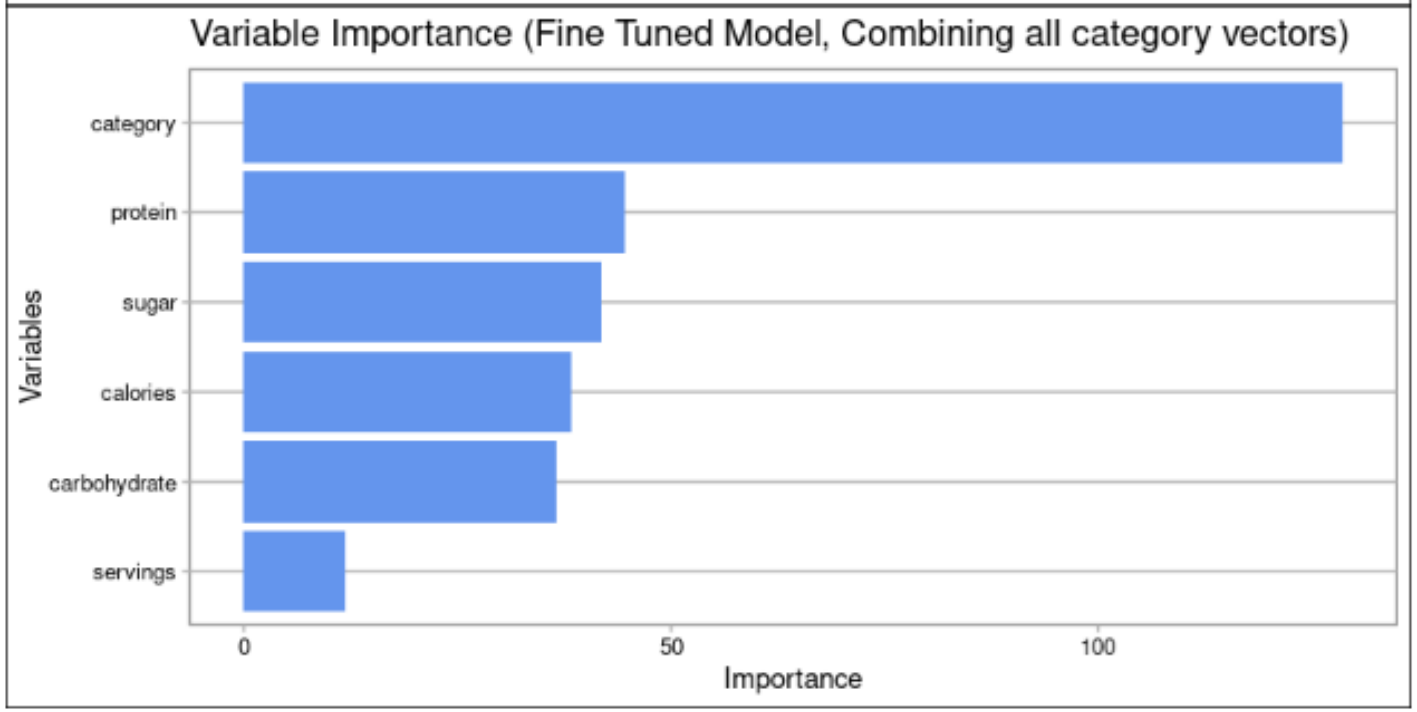


Logistic Regression (LR) vs. Random Forest (RF)

	LR	RF	RF – fine tuned
Accuracy on predicting high traffic	0.76	0.77	0.82
Important features	Category beverages, breakfast, chicken, potato, vegetable, pork, sugar	Category is the most important vector, followed by protein	Category beverages, protein, sugar. Category is the most important vector after combining all the category features

Note: Both the LR and fine-tuned RF models automatically transformed "category" variable into dummy variables.





Vector Importance of the final model

Findings and Recommendations

Findings:

- Recipe category is the most influential factor in determining a recipe's popularity.
- Vegetable, potato, and pork recipes are positively associated with high traffic, while beverage, breakfast, and chicken recipes tend to receive lower traffic.
- The fine-tuned RF model achieved 82% accuracy in predicting high-traffic recipes.

Actions:

- Focus on developing and listing fewer beverage, breakfast, and chicken recipes while expanding vegetable and potato-based recipes.
- Use the fine-tuned RF model to predict a recipe's traffic potential before listing new recipes.

Any Questions?