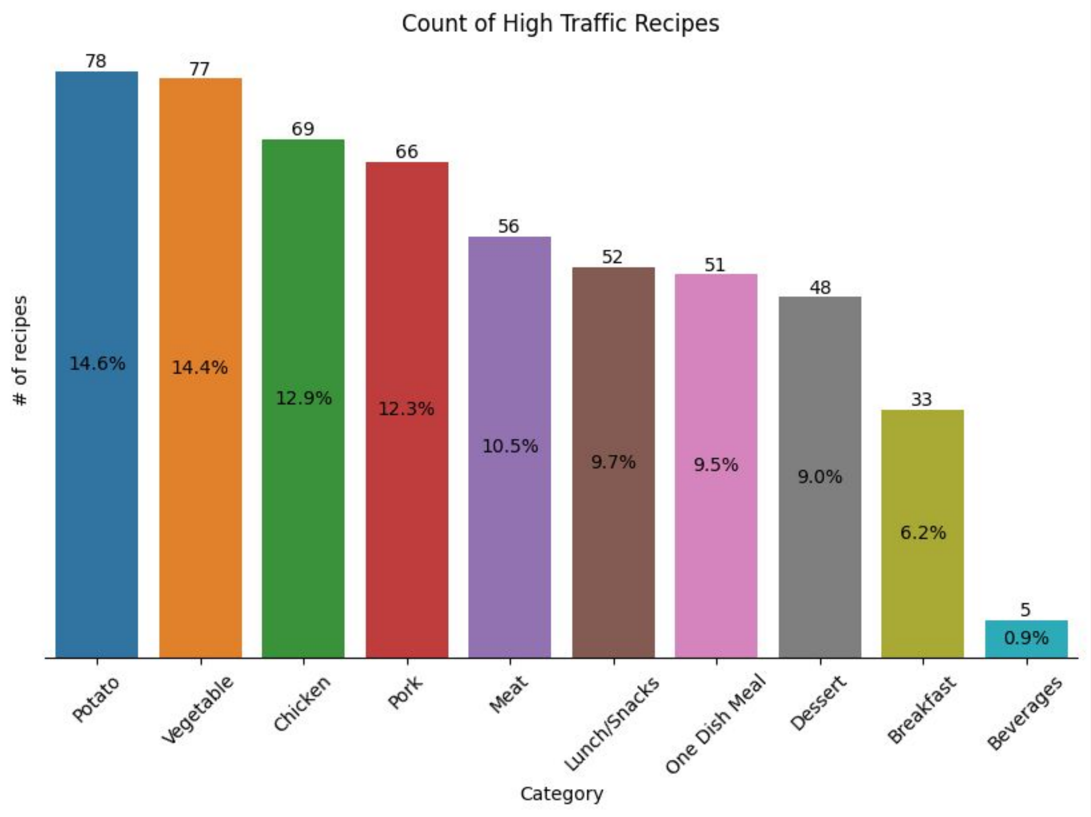
Machine learning models for detecting Recipe Site Traffic

Tasty Bytes

Business Goals:

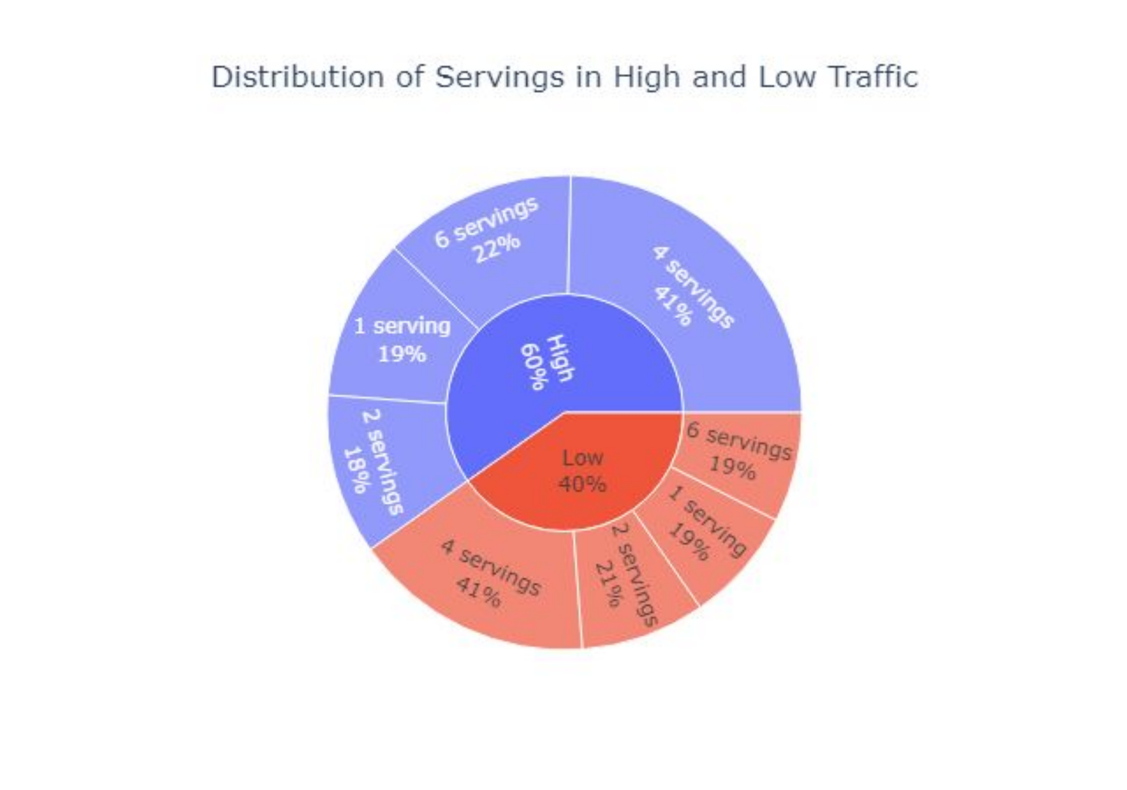
- In a world filled with delicious food recipes, we need to find which ones attract more users
- Our focus is on predicting which recipes will lead to high traffic
- We need to determine a way to predict high traffic recipes 80% of the time

Most Popular type of Recipes



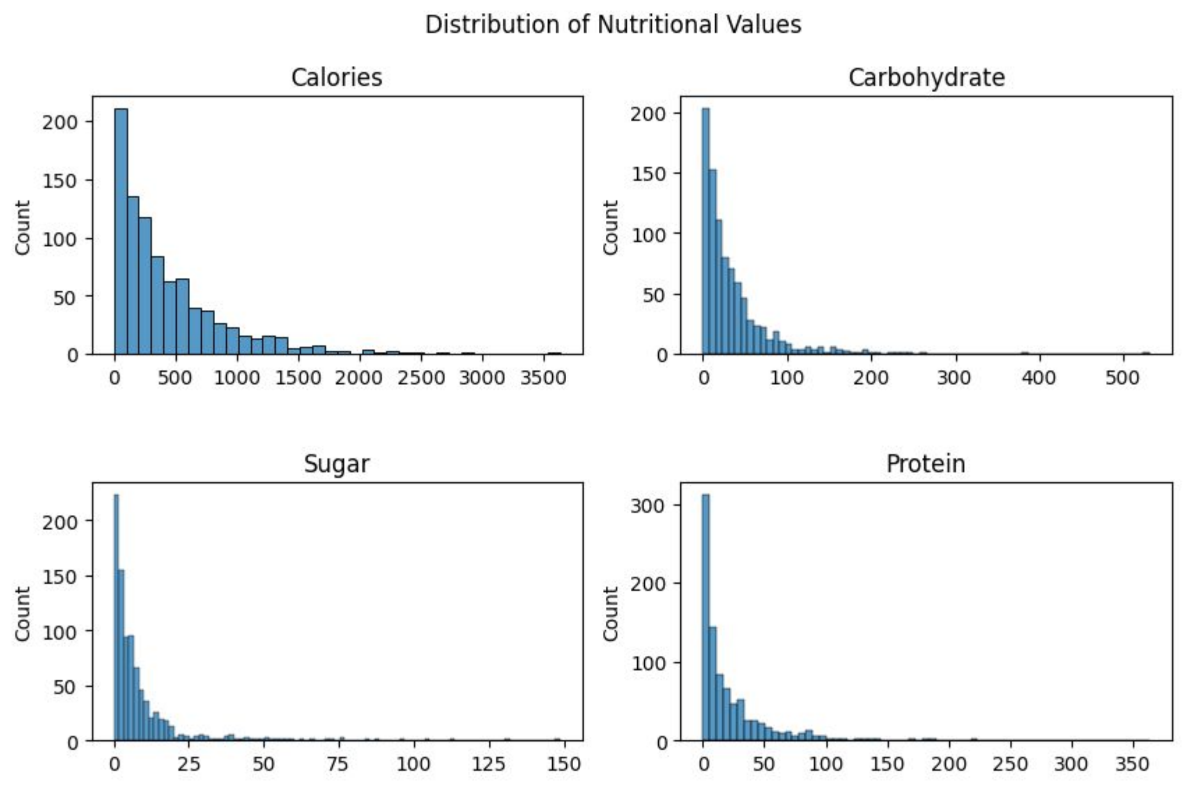
We can observe that our most popular categories of recipes are Potato and Vegetable recipes, which account for 29% of all High traffic. While Breakfast and Drinks receive 6.3% and 0.9% of all High Traffic, respectively, they are our least popular dishes.

Most Popular amount of Servings



We can see that 4 servings, which receive 41% of both High Traffic and Low Traffic, is the most popular number of servings overall. Nonetheless, the remaining servings alternatives generate almost the same amount in both groups.

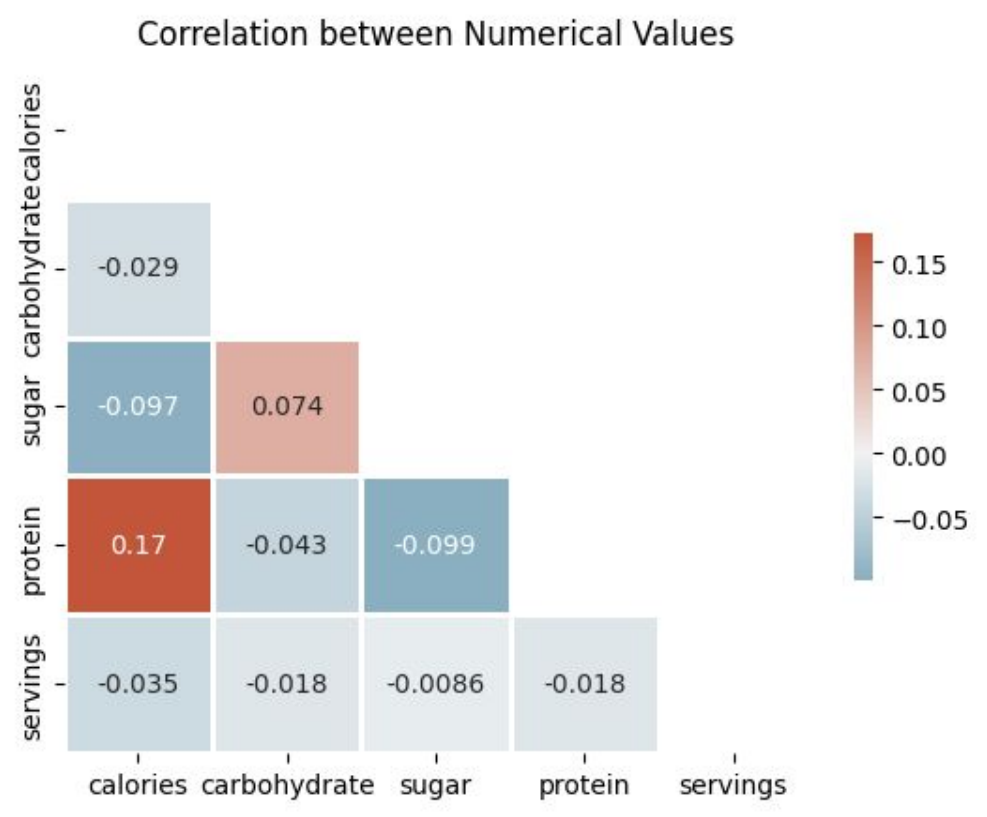
Spread of Nutritional Values



We can see that the distribution of our Nutritional Values columns is all right skewed, which means that a few recipes are really high in nutritional values.

These values must be standardised when creating a model.

Correlation between Columns



When we examine the Correlation between our numerical values, we find that there is no significant pearson correlation between any 2 variables, with proteins and calories having the largest correlation (0.17).

Model Development:

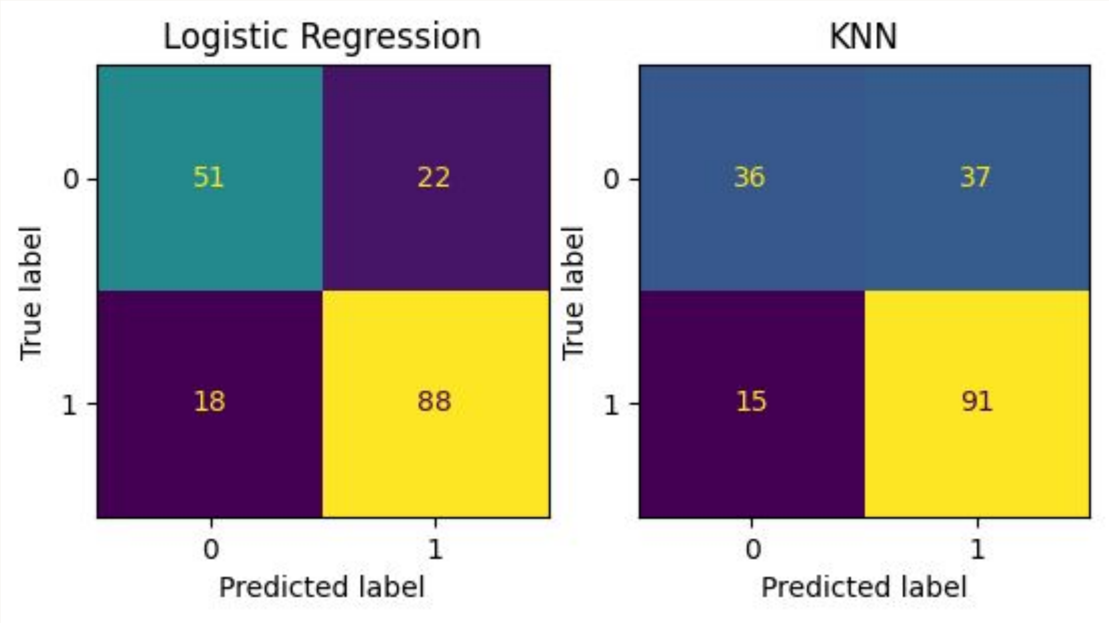
This is a Classification challenge because we're determining whether or not a recipe receives a lot of traffic.

The first step is to divide the data into input and output variables. Then, we encode the categorical data and scale the numerical data. Next, we divided the data into training and testing groups of 80/20.

Our basic model is a KNN model and our comparison model is a logistic regression model.

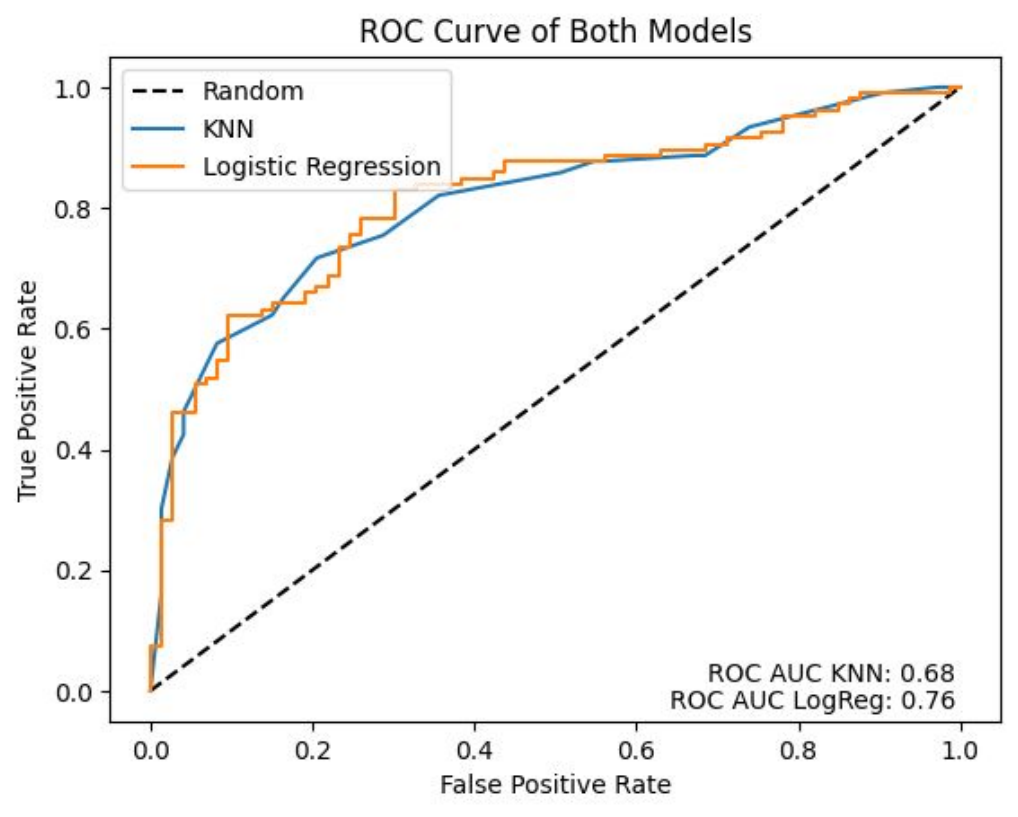
Both use GridSearch to find optimal hyperparameters.

Confusion Matrix comparison



We can see that the Logistic Regression is either performing similarly or better than the KNN model

ROC Curve of Both models



The Logistic Regression model has a greater true positive rate and a lower false positive rate when compared to the KNN model, according to the ROC curves of both models.

We can also see that the ROC AUC score of the Logistic Regression model is higher than the KNN, confirming that the LogReg model is better for detecting High traffic recipes

Business Metrics

Measuring Precision, Recall, and F1-score is an effective metric to evaluate the accomplishment of our two goals.

While The accuracy score and ROC AUC score are useful KPIs for these models to gauge their predictive power.

Our logistic regression model performs better than the KNN model in all the mentioned metrics

Recommendation

We can put our logistic regression model into production to aid the product manager in predicting the high traffic of the recipes. By using this model, approximately 80% of the predictions will ensure that there will be heavy traffic.

To implement and improve the model, I will consider the following steps:

1. Deploying the model into production, preferably using edge devices for ease and security
2. Collecting more data such as time to make, ingredients, site duration time,etc.
3. Feature Engineering such as increasing the number of values in category column, creating more meaningful features from the variables

Thank You