Predicting High-Traffic Recipes

Maximizing Website Engagement

Ezzaddeen Mofarreh

Agenda

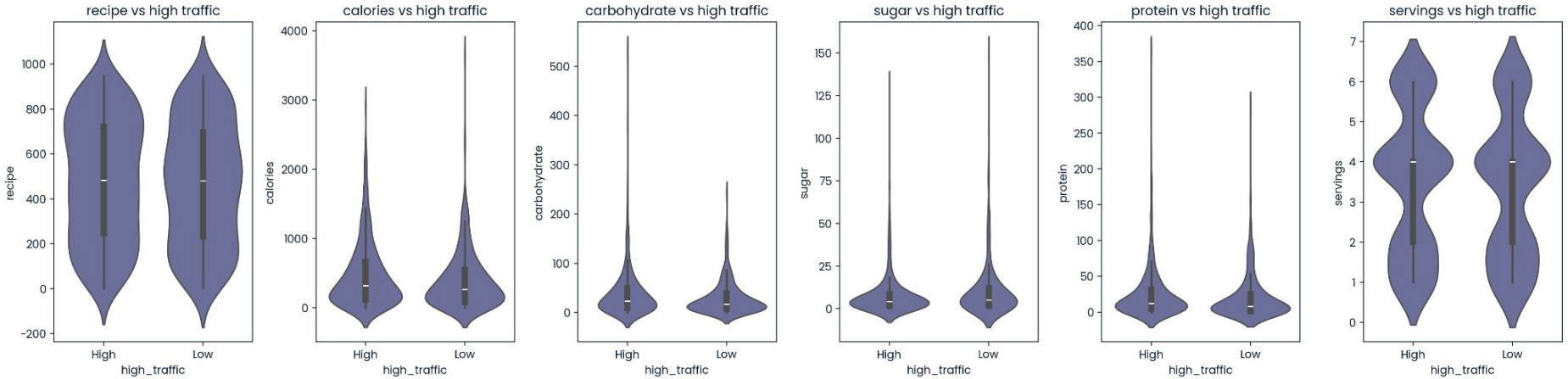
- Business Objectives
- Predictive Power of Numeric Features for Traffic Levels
- Recipe Category Distribution
- Category Effectiveness for Generating High Traffic
- Model Development:
  - Evaluating LinearSVC Performance and Feature Importance
  - Evaluating Naive Bayes - GaussianNB - Performance and Feature Importance
- Models Comparison
- Business Monitoring Metric
- Recommendations

Business Objective

- Predicting **which recipes will lead to high traffic** on the homepage to increase traffic and subscriptions, with a target **precision of 80%**. (Correctly predict high traffic recipes 80% of the time)

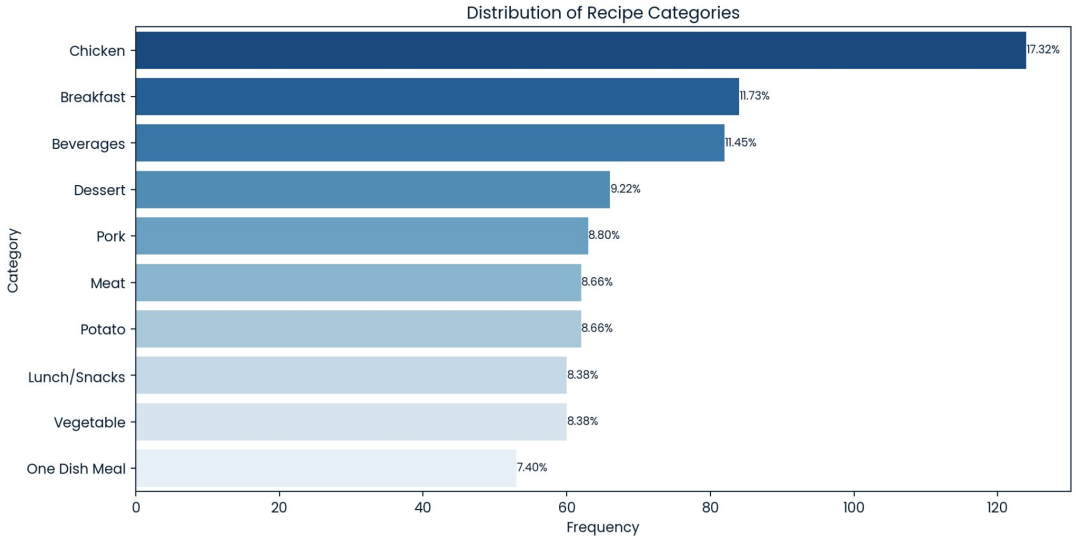
- **Traffic goes up by as much as 40%** to the rest of the website if a popular recipe is chosen. More traffic means more subscriptions so this is really important to the company.

Predictive Power of Numeric Features for Traffic Levels



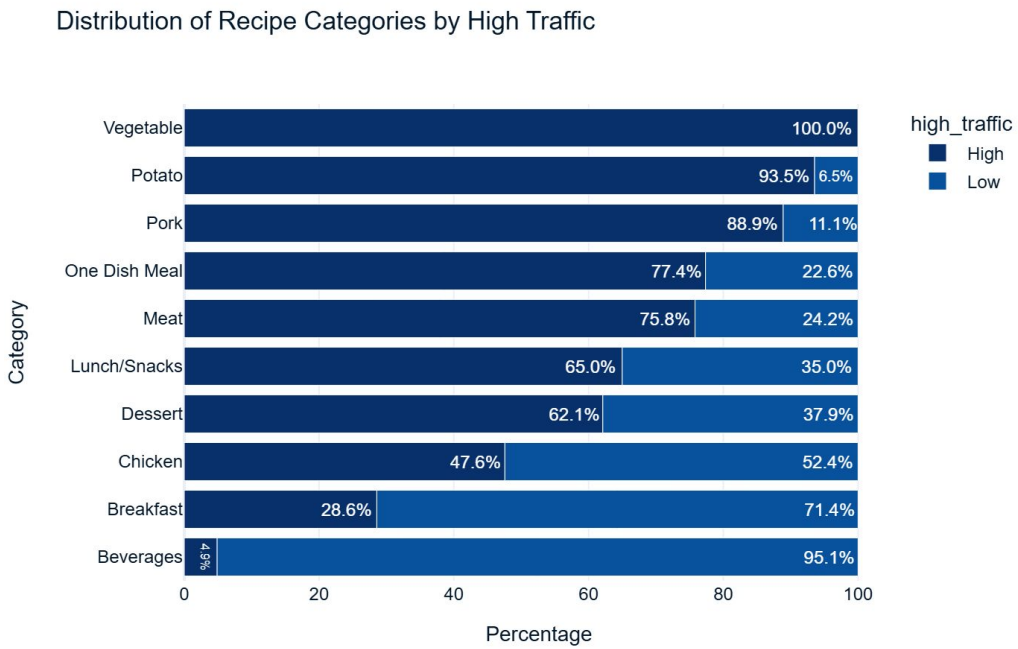
The overall distributions of the numeric features (calories, carbohydrates, sugar, protein, and servings) appear similar between high and low traffic categories. This suggests that **these features alone may not be strong predictors of traffic levels**.

Recipe Category Distribution



- The distribution of recipe categories shows that **Chicken is the most common**, comprising approximately 17.3% of the dataset.
- Breakfast and Beverages follow, each making up around 11.7% and 11.5%, respectively.
- Other categories have similar percentages.
- **One Dish Meal recipes are the least common**, representing 7.4% of the total.

Category Effectiveness for Generating High Traffic

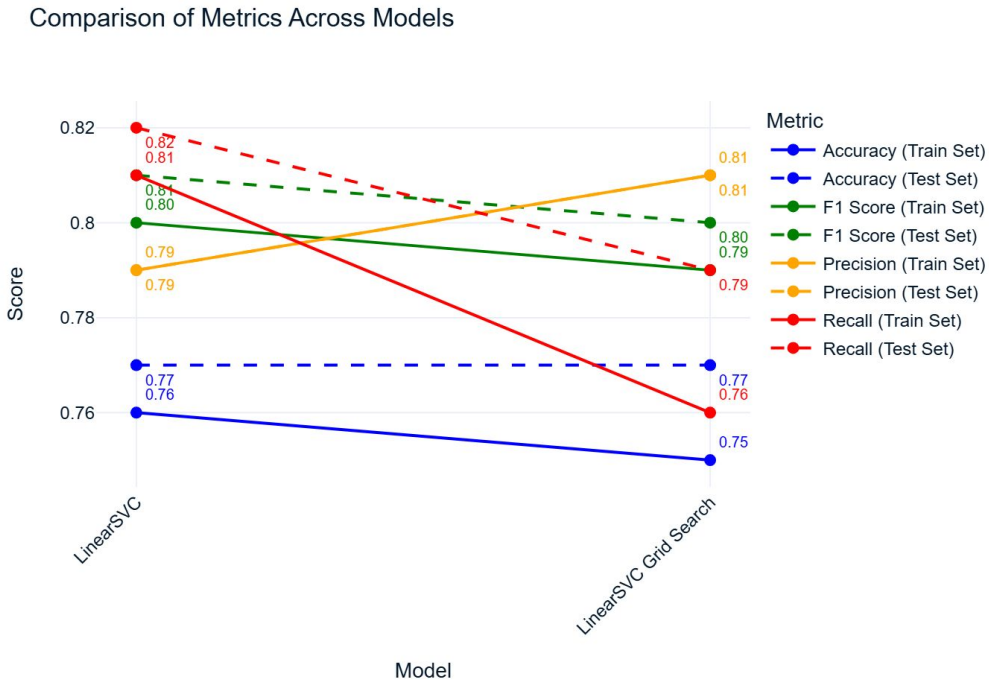


- **'Vegetable,' 'Potato,' and 'Pork'** are **particularly effective** at generating high traffic and should be prioritized when featuring recipes on the homepage.
- **'One Dish Meal' and 'Meat'** also **show potential** for high traffic.
- **'Lunch/Snacks,' 'Dessert,' and 'Chicken'** are **moderately successful** and can be included with a balanced approach.
- **'Breakfast' and 'Beverages'** are **less effective** in driving high traffic and may require less frequent featuring or strategic pairing to enhance their impact.

Precision vs Recall

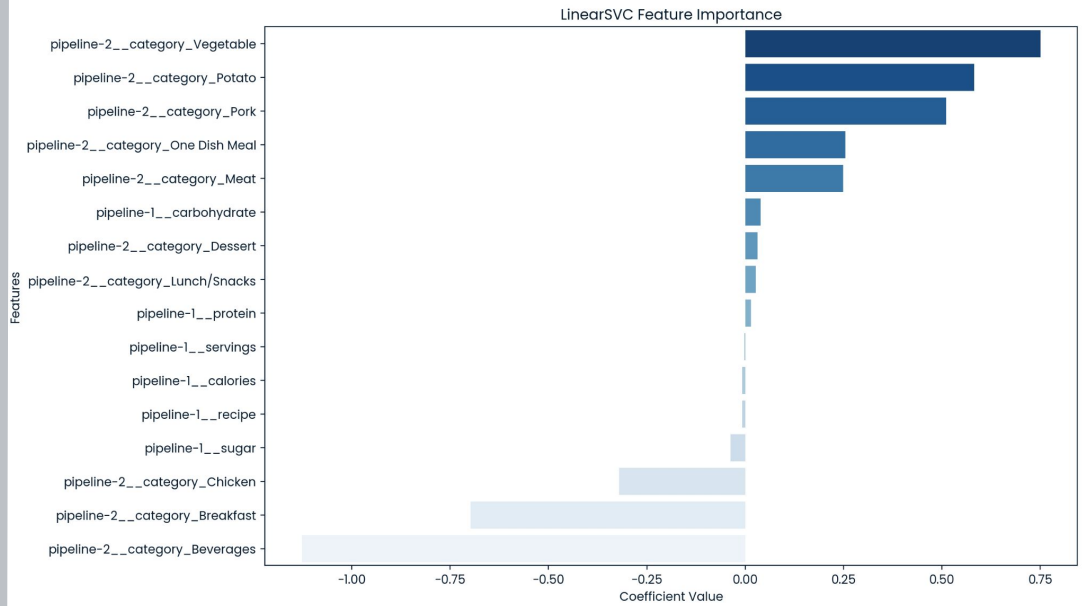
- **Precision** is the metric that directly measures the proportion of correctly identified high traffic recipes among all recipes predicted as high traffic. The business target is a precision of 80%, meaning that at least 80% of **the recipes predicted to have high traffic should indeed have high traffic**.
- **Recall** measures **the proportion of actual high traffic recipes correctly identified** by the model. It helps in understanding how well the model captures all potential high traffic recipes.

LinearSVC Model Performance



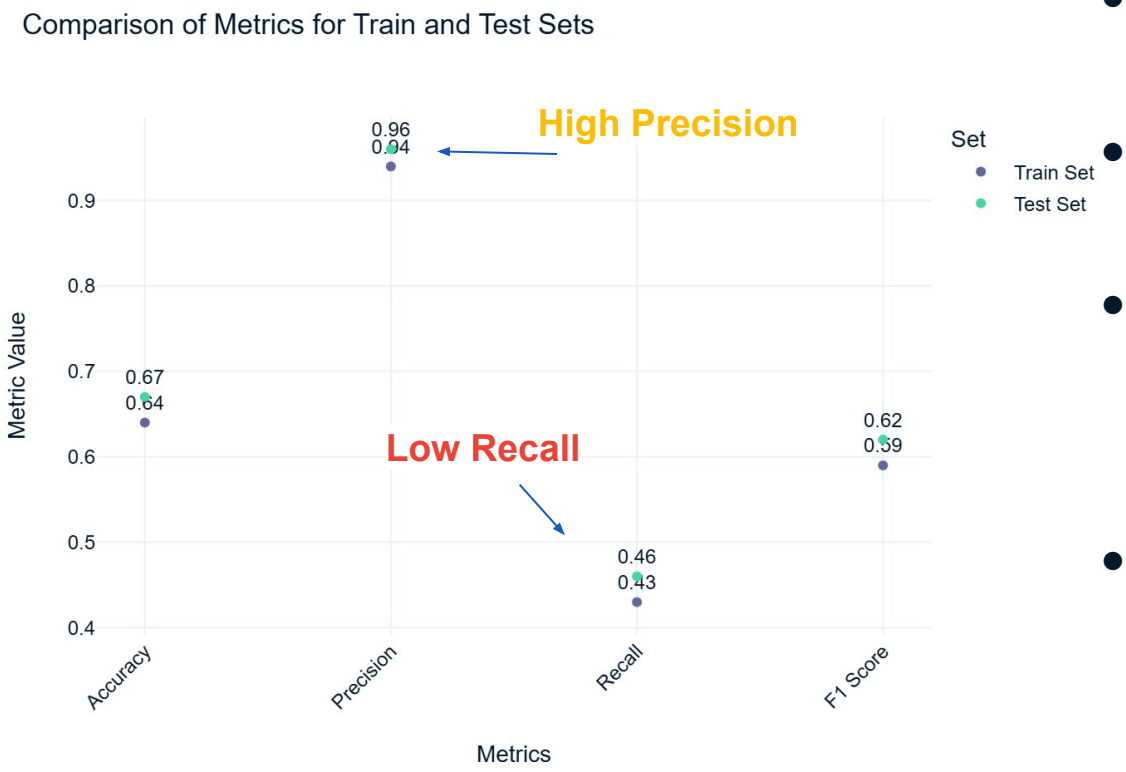
- Both models exhibit **strong generalization**, with consistent metrics across training and test sets.
- The LinearSVC Grid Search model shows a slight improvement in precision, **surpassing 80%**, compared to the base model.
- Both models effectively **balance precision and recall**.

LinearSVC Feature Importance



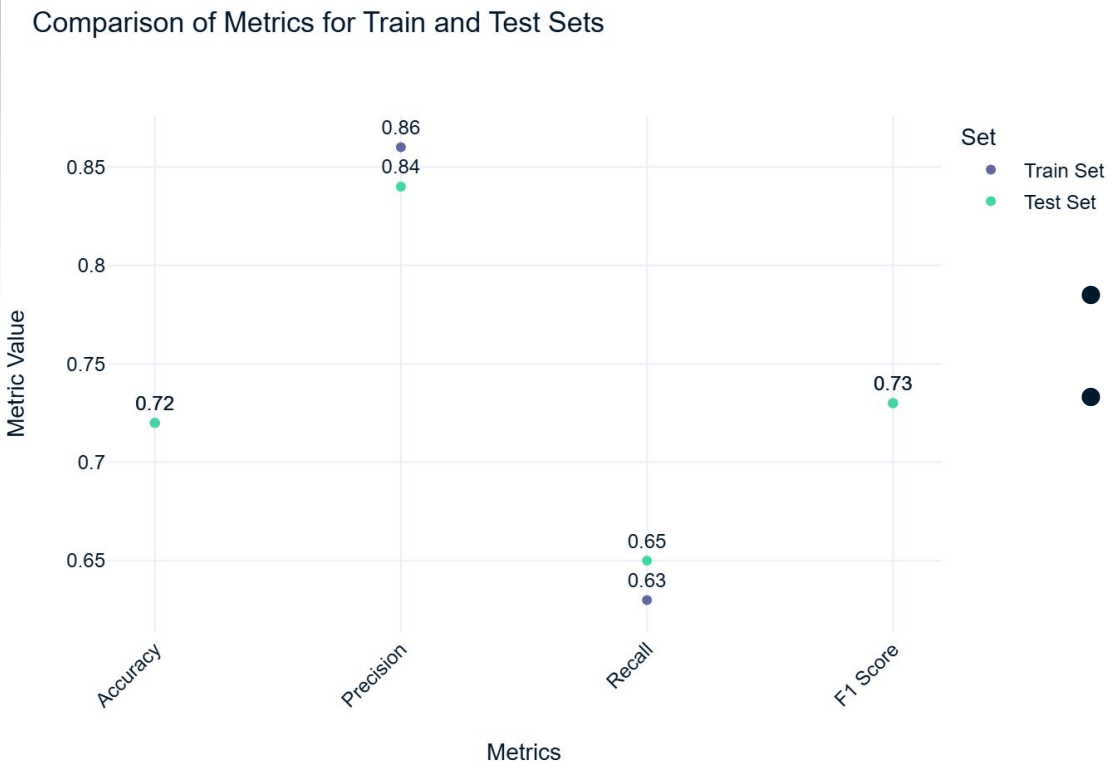
- **'Vegetable,' 'Potato,' and 'Pork'** which identified as particularly effective at generating high traffic, are among **the top features** the model has learned. These categories should be prominently featured to maximize traffic.
- **'Breakfast' and 'Beverages'** which have been found to be less effective at driving high traffic, are among the **least impactful features**.
- Feature importances align well with the observed high traffic proportions for different categories.

Naive Bayes Base Model Performance



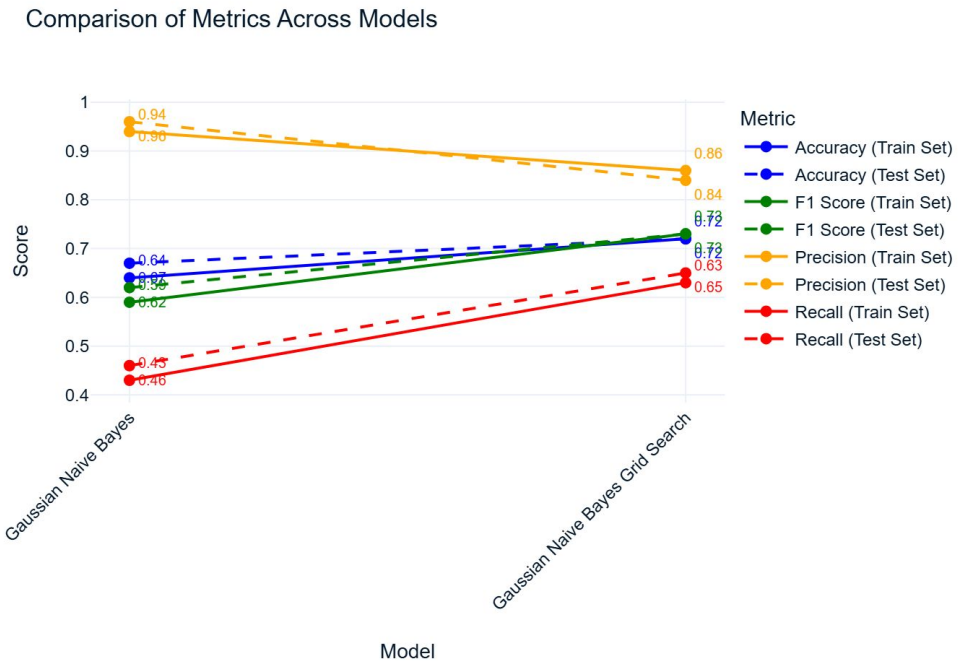
- Model performance is **stable and consistent** across the train and test set evaluations.
- However, the model does exhibit a **trade-off between precision and recall**
- The model consistently demonstrates **high precision**, which is crucial for ensuring that the recipes predicted to be popular are indeed likely to drive traffic.
- The model consistently demonstrates **low recall** indicating that the model is not identifying a large portion of actual popular recipes.

Naive Bayes Tuned Model Performance



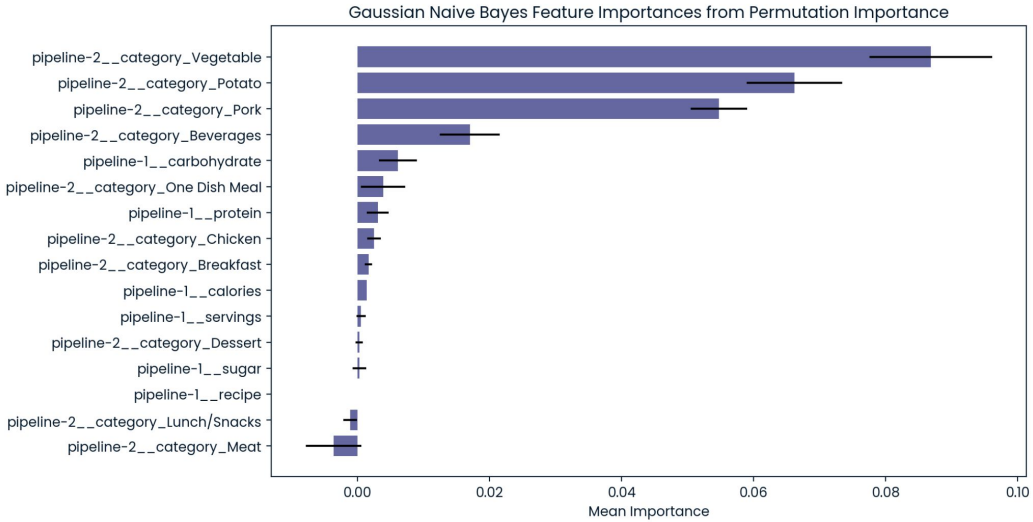
- **Precision remains high**, indicating reliable positive predictions.
- **Recall has improved** compared to the base model, showing that the model now identifies a higher proportion of actual positive cases.

Naive Bayes Model Performance



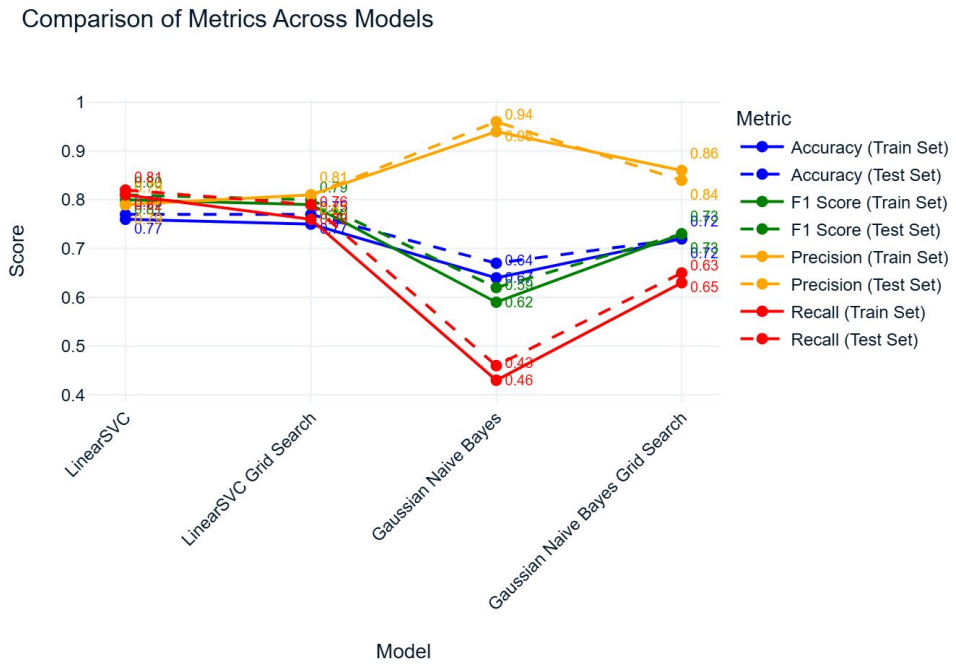
- Both models exhibit **strong generalization**, with consistent metrics across training and test sets.
- The base model shows **high precision but lower recall**, indicating that while it is very accurate in its positive predictions, it misses a significant portion of actual positive cases.
- The Grid Search-optimized model shows significant improvements, **precision remains high, and recall has improved**, reflecting the model's enhanced ability to identify positive cases.

Naive Bayes Feature Importance



- **Top Features**: 'Vegetable', 'Potato', and 'Pork' categories are the most influential in the model, **consistent with the LinearSVC model's findings.**
- **Moderate Features**: 'Beverages' and 'One Dish Meal' play a moderate role.
- **Least Important Features**: 'recipe', 'sugar', and 'Dessert' categories have minimal impact.
- **Negative Impact Features**: 'Lunch/Snacks' and 'Meat' negatively affect the model's performance.

Models Comparison



- **LinearSVC Models**: Both the base and grid search models **perform consistently well**, with grid search making only minor adjustments that do not significantly impact overall performance.
- **Gaussian Naive Bayes Models:** The base model shows high precision but low recall, resulting in imbalanced performance. The grid search-optimized model addresses this imbalance, achieving **improved precision and recall, and demonstrating better overall performance** on both training and test sets.

Business Monitoring Metric

To effectively monitor the achievement of **the business objective (predicting high traffic recipes)**, the following metrics should be tracked::

- **Precision**: This is **the primary metric** for monitoring as it directly measures the proportion of correctly identified high traffic recipes among all recipes predicted as high traffic. The business target is a precision of 80%, meaning that at least 80% of the recipes predicted to have high traffic should indeed have high traffic.
- **Recall, F1 Score, and Accuracy**: While precision is the main focus, these metrics provide additional insights into the model's overall performance. Recall measures the proportion of actual high-traffic recipes identified by the model, F1 Score balances precision and recall, and accuracy provides a general view of model performance.
- **Traffic Impact**: Track the increase in website traffic resulting from the model's recommendations. This is a direct measure of how well the model contributes to achieving the business goal.

Recommendations

- **Monitor Precision Closely**
  To achieve the business objective of increasing traffic and subscriptions, monitor precision closely, **aiming to meet or exceed the 80% target**. Also, keep track of recall, accuracy, and the impact on website traffic to ensure that the model contributes effectively to the business goals

- **Adopt the LinearSVC Grid Search Model**
  The Grid Search-tuned LinearSVC model is recommended for production due to its balanced performance, with consistent results across both datasets and high precision. It aligns well with the business objective of predicting high traffic recipes with at least 80% precision. This model meets the precision target with **a precision score of 0.81** on both the training and test sets, demonstrating its effectiveness at accurately identifying high-traffic recipes.

THANK YOU