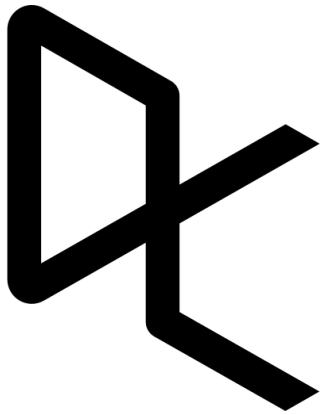


Development of a model that correctly predict High traffic recipes

Presented by Mohamed Shehata

Professional data scientist Data camp Certification Exam

Project objective

- Predict which recipes will lead to high traffic.

- Correctly predict high traffic recipes 80% of the time.

Data Validation

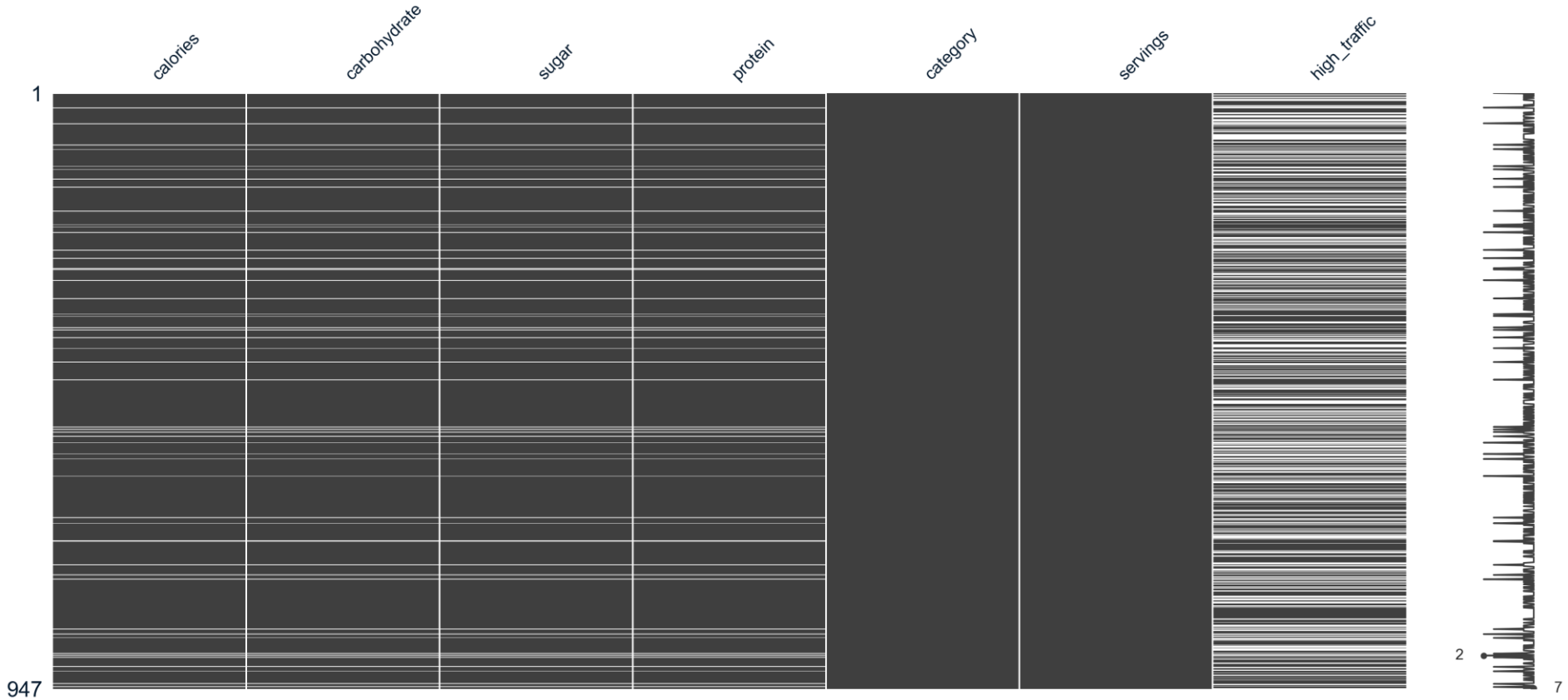


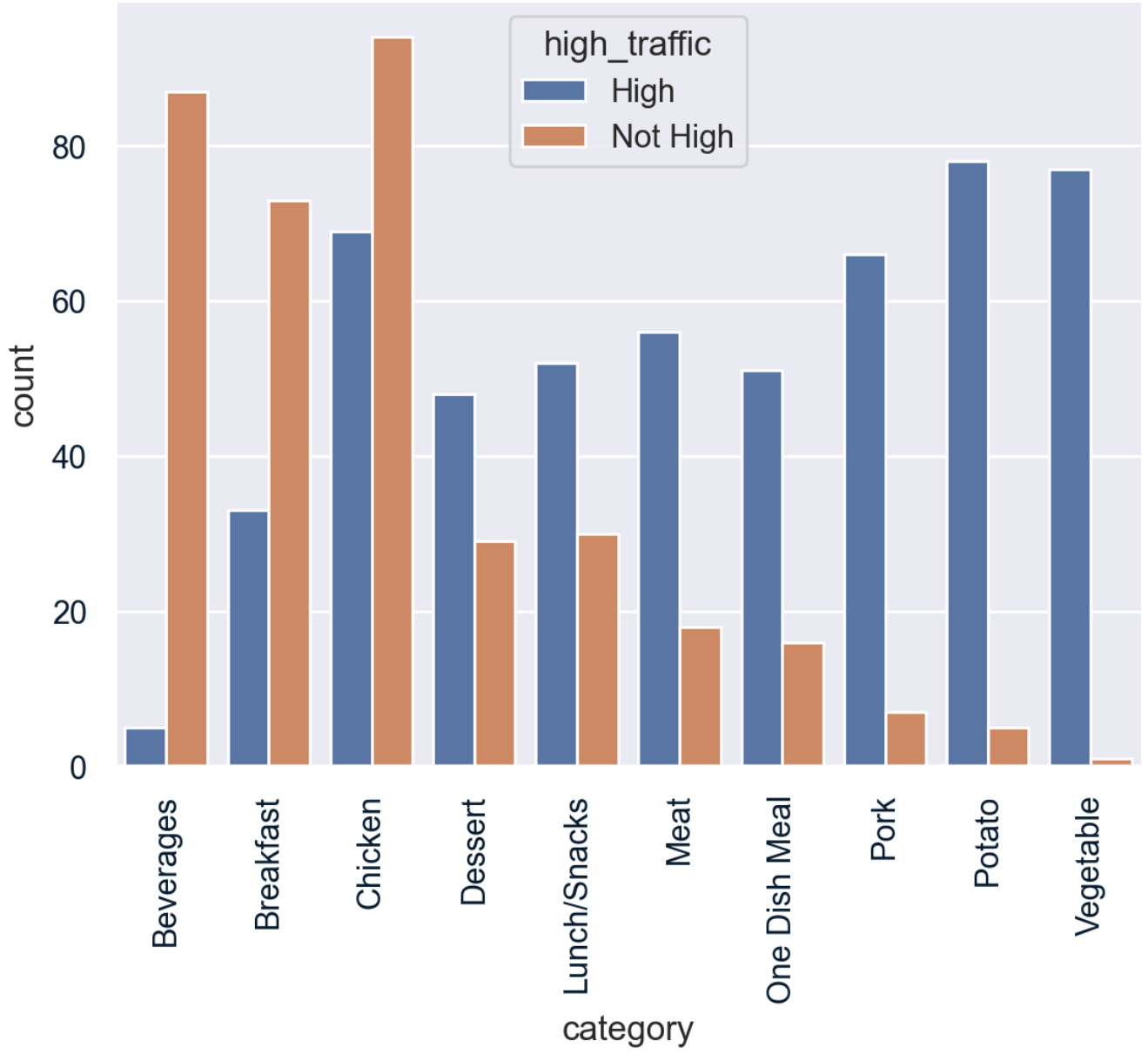
Diagram shows white strips as NaN or empty data

Data Validation

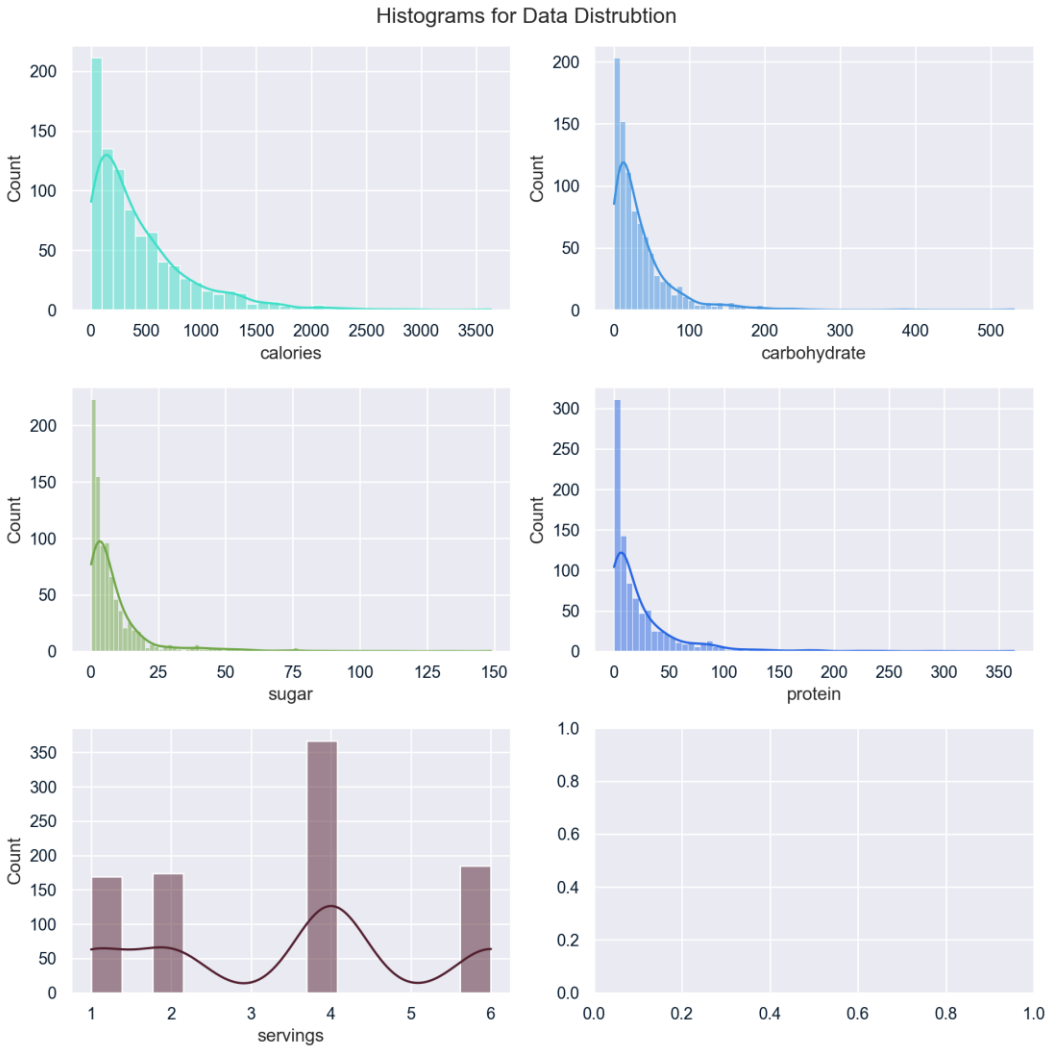
- **recipe:** set as an index for the data frame since its unique identifier
- **calories:** no changes have been applied
- **carbohydrate:** no changes have been applied
- **sugar:** no changes have been applied
- **protein:** no changes have been applied
- **category**: chicken breast was renamed to fit the chicken category and the type of the column was changed to category
- **servings**: removed redundant strings that conveyed the category since there is already a category column and the type of column was changed to int
- **high_traffic**: Since the team documentation showed that only high-traffic recipes were labeled the NAN values were replaced with 'not high' string
- **Additionally,** NAN values of other columns were removed changing the no of recipes from 947 to 895 which is considered an acceptable amount of data to proceed with the analysis and modeling

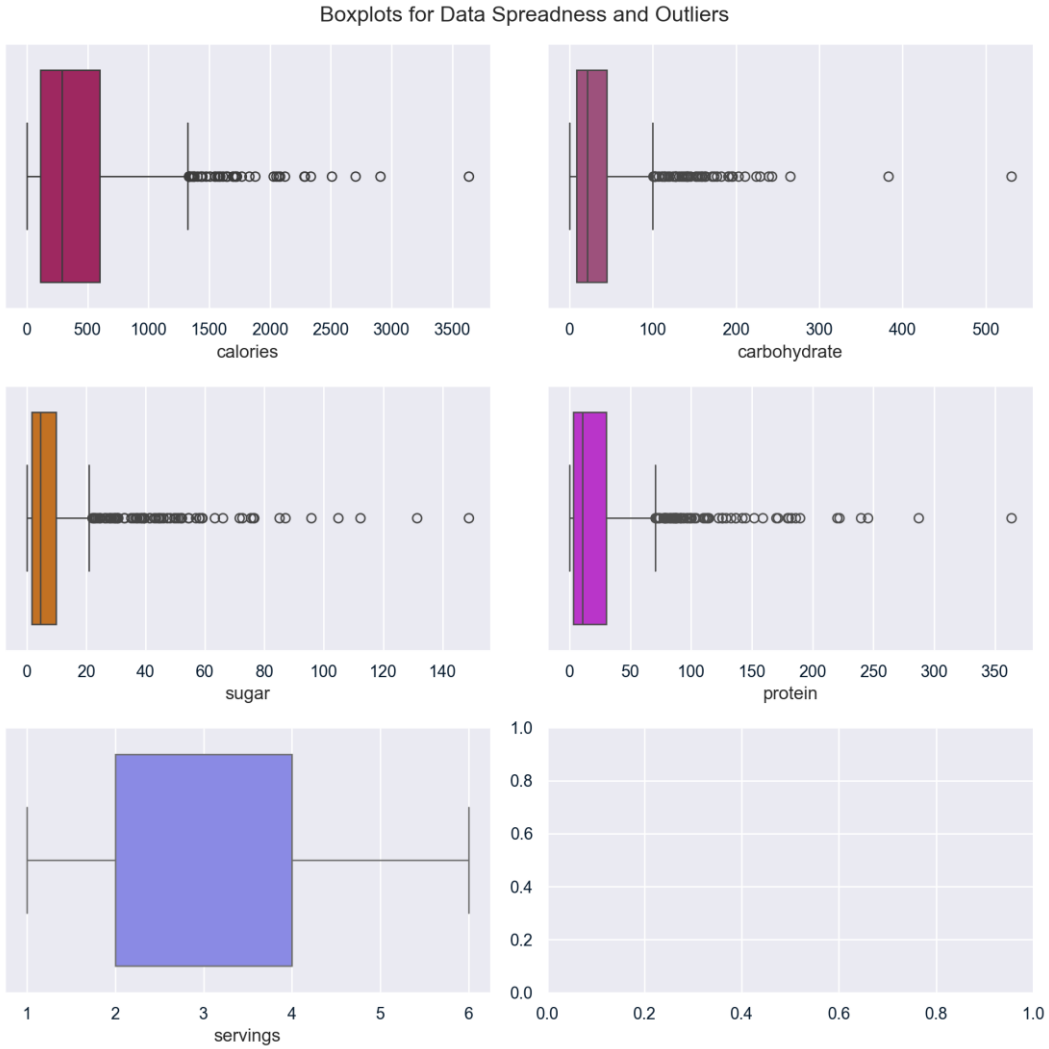
Exploratory Data Analysis

According to this graph, we can see clearly that vegetables, pork, and potatoes have a very high ratio of being rated high traffic with a low ratio not being high. while chicken seems to be the 3rd highest in high traffic rate, it also has a very high ratio to low traffic. beverages seem to be the least high-traffic category



Exploratory Data Analysis

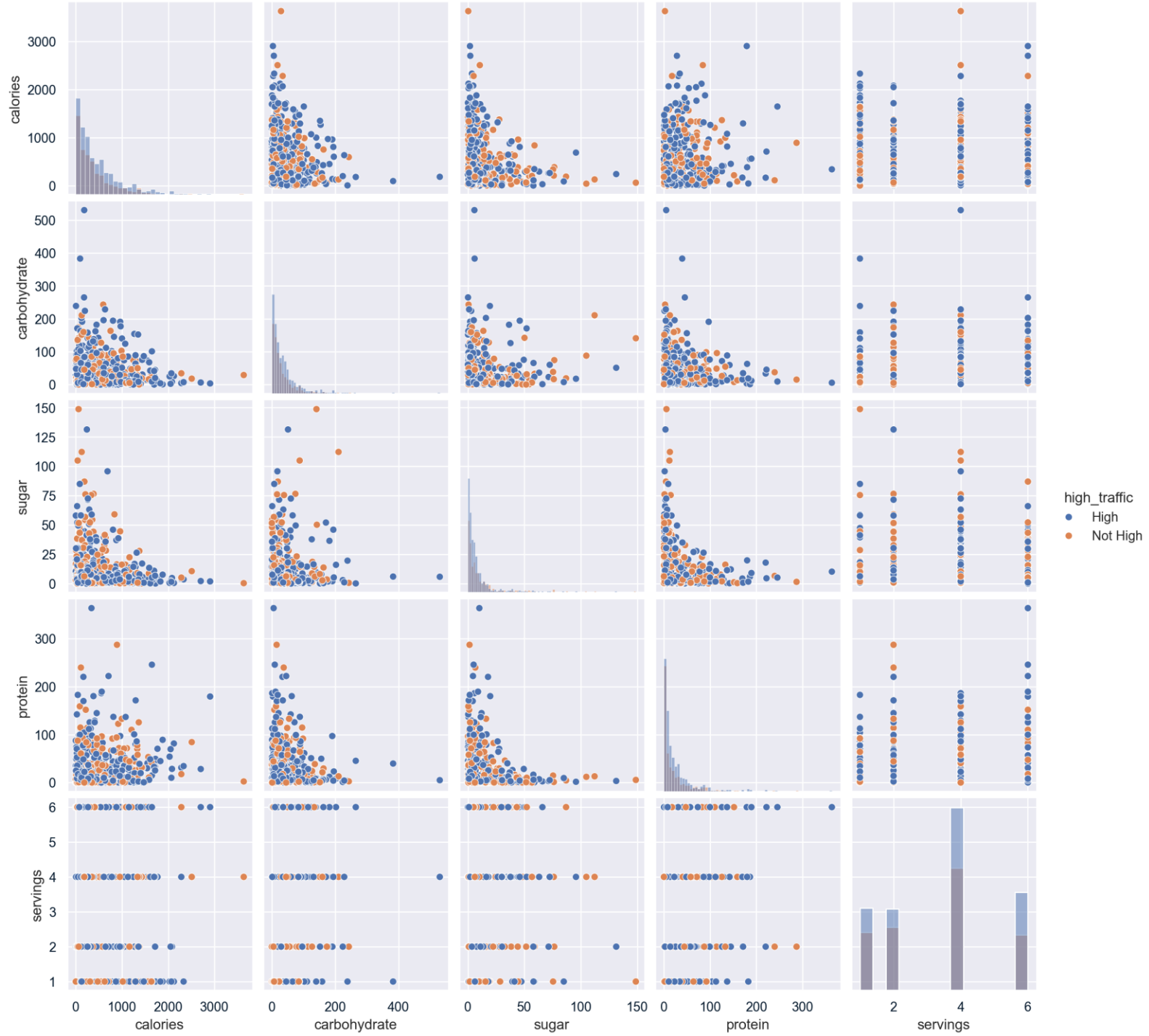




Graphs to investigate distribution, outlier or data spread

Exploratory Data Analysis

Pair plot to investigate Correlations



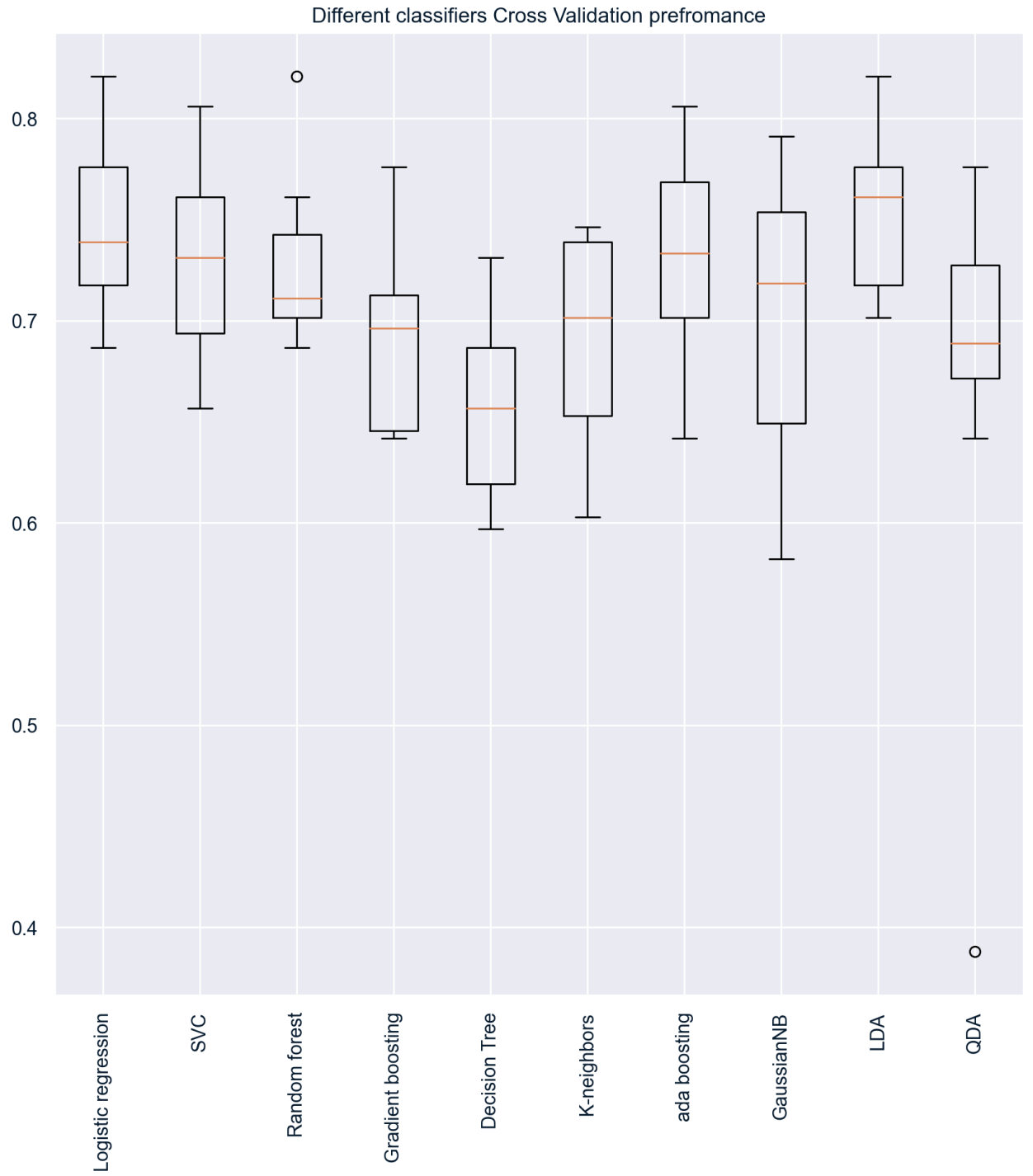
Models development

The feature columns were modified to prevent any bias and proper data representation for the models. The engineered features are;

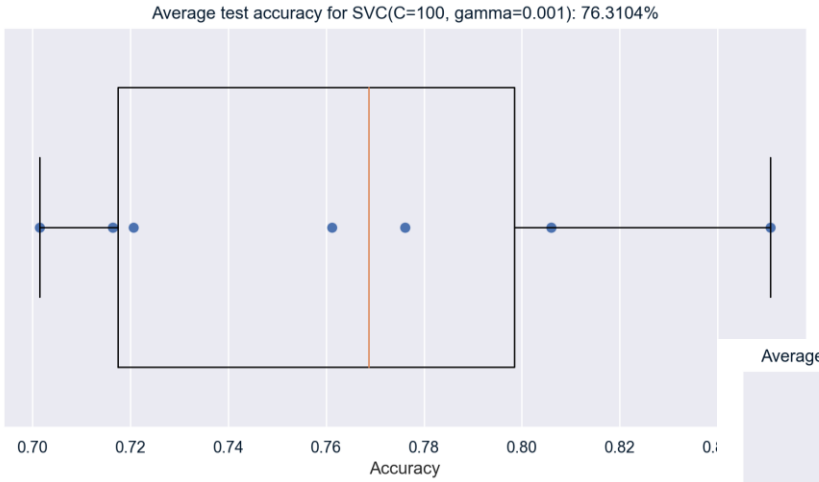
1. Applied yeo-johnson power transformation -(not Box-Cox because it deal with only positive values)- for the numeric columns to reduce skewness
2. Encoded nominal Categorical data (category column)
3. Mapped ordinal categorical data (High_traffic column)
4. Split the data for validation and ensure stratification of the split
5. Defined Kfolds for Cross-validation of the models
6. Investigated the existence of collinearity after feature engineering to ensure proper feature selection
7. Decide on which models perform well on our data frame

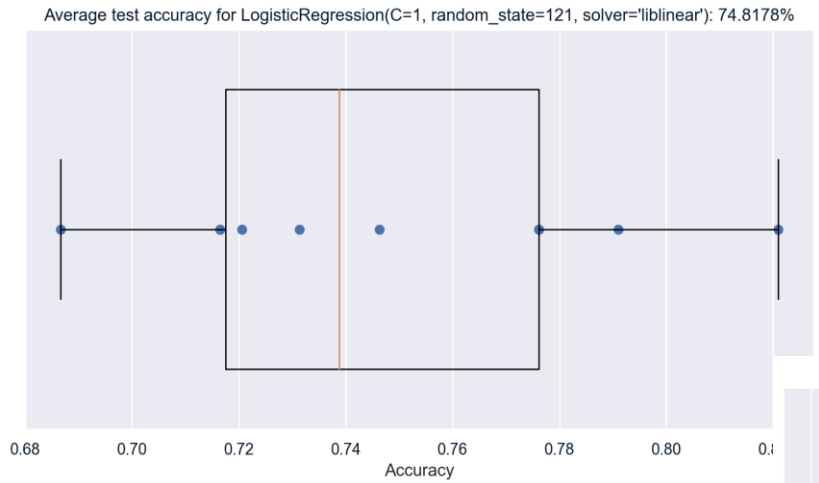
Models development

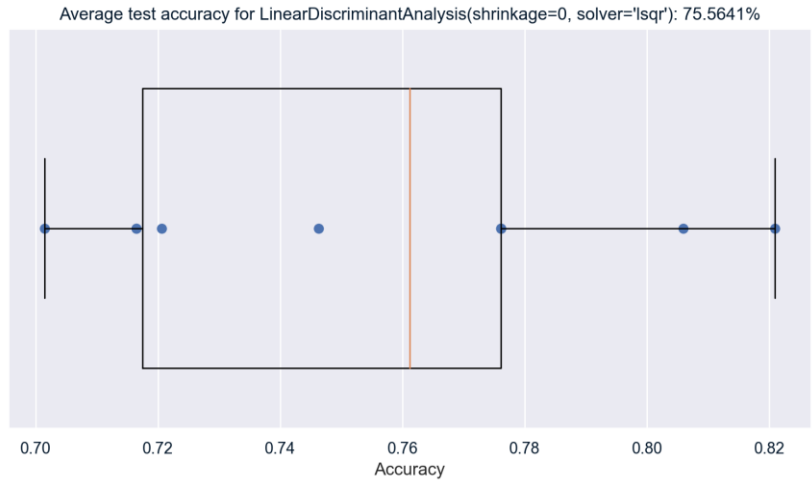
Logistic regression
f1 score = 0.8116 MSE = 0.2321
SVC
f1 score = 0.7986 MSE = 0.2589
Random forest
f1 score = 0.7746 MSE = 0.2857
Gradient boosting
f1 score = 0.7536 MSE = 0.3036
Decision Tree
f1 score = 0.7306 MSE = 0.3259
K-neighbors
f1 score = 0.7491 MSE = 0.3080
ada boosting
f1 score = 0.7891 MSE = 0.2768
GaussianNB
f1 score = 0.7243 MSE = 0.2991
LDA
f1 score = 0.8015 MSE = 0.2411
QDA
f1 score = 0.7510 MSE = 0.2812



Models Evaluation







Box plots for cross-validation

Models Evaluation

- Logistic Regression Model F1_score = 0.8116 meets company goal
- Linear Discriminant Analysis Model F1_score = 0.8015 meets company goal
- SVC Model F1_score = 0.8015 meets company goal

Conclusion and Recommendations

The aim of the project is to develop a model that can correctly predict high-traffic recipes 80% at a time. the dataset was, manipulated and analyzed to draw helpful conclusions in feature engineering of our dataset followed by a comparison of different sklearn classifiers. logistics model, Support vector classifier, and linear discrimination analysis were shown to be the best-performing classifiers. upon the fine-tuning of the parameters and evaluation, we can conclude that for our current scenario, logistic the model is the best choice in terms of our goal and low MSE.

1. **Although bootstrapping was shown to not be significantly effective, more datasets may improve the performance if it was of different quality**
2. **Investigate further why vegetables as the most high-traffic category by testing of recipes containing more vegetables will lead to more probable high traffic (Hypothesis testing)**
3. **Add the factor of time in our data since it may draw new correlations and meaningful insights to increase the tendency of the model to predict high-traffic**

Thank you for your time