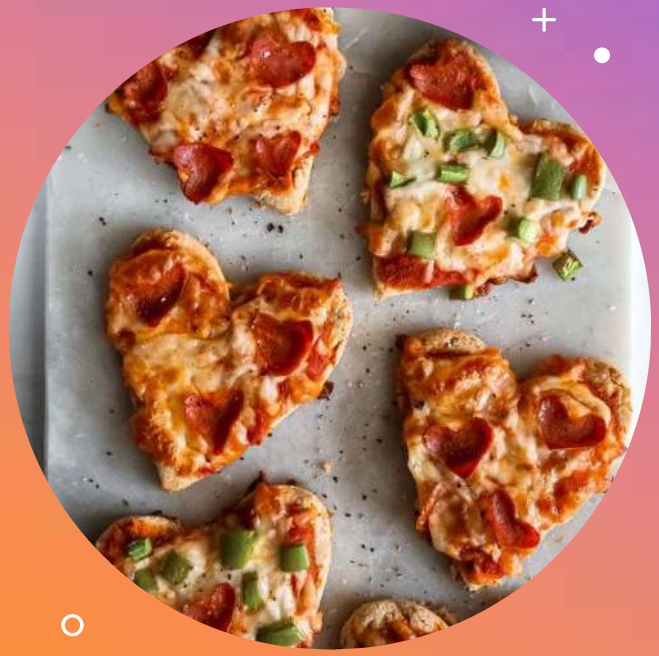
RECIPE WITH HIGH TRAFFIC

Hossam Ahmed Salah



AGENDA

Business objective

KPI

Data Validation

EDA

Models and Business metrics

What model satisfy the KPI?

Recommendation for the future

Tasty Bytes

- Tasty Bytes was founded in 2020 during the Covid Pandemic.
- It started as a recipe search engine to help people utilize limited supplies.
- It has evolved into a fully-established business.
- Tasty Bytes offers a monthly subscription for a comprehensive meal plan.
- The meal plan ensures a healthy and well-balanced diet.
- Subscribers can have the ingredients delivered to their doorstep.
- Tasty Bytes aims to provide **inspiration** and support during challenging times.

BUSINESS OBJECTIVE

Objectives

- Providing inspiration for everyone as Tasty Bytes considered a pioneer in this field.
- Increasing the Profit for the company.
  - By increasing the subscriptions
  - For more subscriptions we need more traffic
  - Traffic high when we have popular recipes in the homepage of the site
  - Traffic to the rest of the website increase 40% in the case of popular recipes
  - We need to predict recipes that would be popular so lead to high traffic
  - The accuracy that the business can rely on it can't be less than 80% most of the time
  - 80% correctly and reliably predict Hight traffic recipes
  - Recommendation for the business about what features make the recipes popular so they can focus on publishing more of these recipes
  - Recommendation for any improvement over the business flow and design based on the analysis

KEY PERFORMANCE INDICATORS

KPI(s)

- **Increase the traffic 40%**
- **Predict which recipes will be popular 80% of the time**

DATA VALIDATION

Data Validation

- **Dataset has 947 rows, 8 columns**
- **Recipe ☑ Numerical id**
- **Calories ✗ Numerical has missing data**
- **Carbohydrates ✗ Numerical has missing data**
- **Sugar ✗ Numerical has missing data**
- **Protein ✗ Numerical has missing data**
- **Category ✗ String type but it 11 categories not 10 has no missing data**
- **Servings ✗ Not Numerical as some values has a text like ‘4 as snack’ no missing data**
- **High-traffic ☑ the missing data represent the recipes with low traffic**

Data Validation

Dataset has 947 rows, 8 columns

For missing data we choose to fill it with median for the numerical features

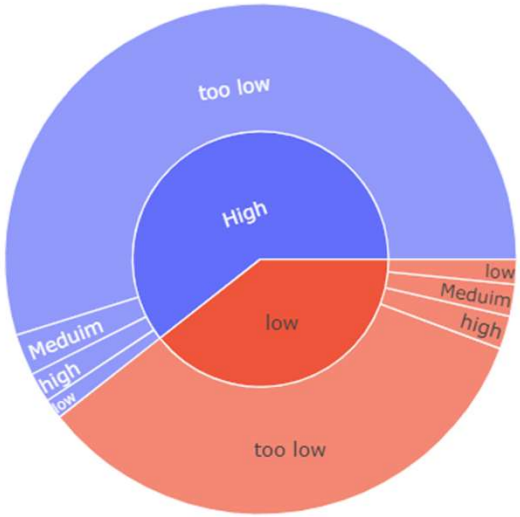
For missing data in the target we fill it with 'low' to be other class we have in this problem 'A Binary Classification'

I did Feature Extraction to extract more insights from the data by making categories out of the numerical data

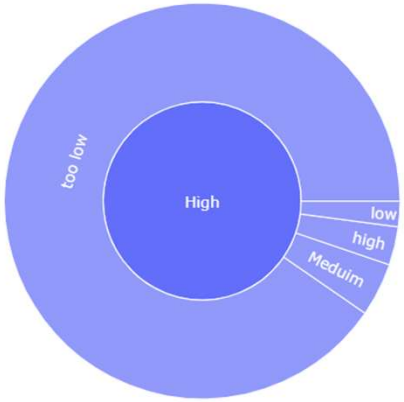
To prepare the data for the model we should make all the columns numerical this step is the 'Encoding step' and later we would scale the data

EXPLORATORY DATA ANALYSIS

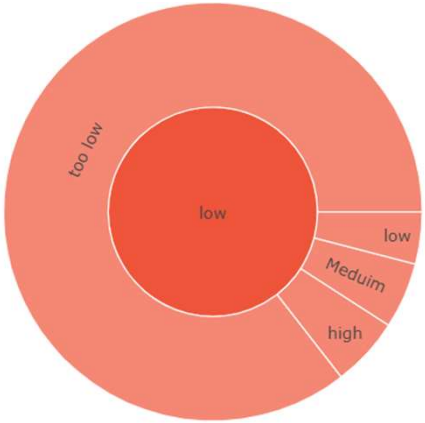
Traffic Vs Sugar categories



- High traffic 574
- Low traffic 373

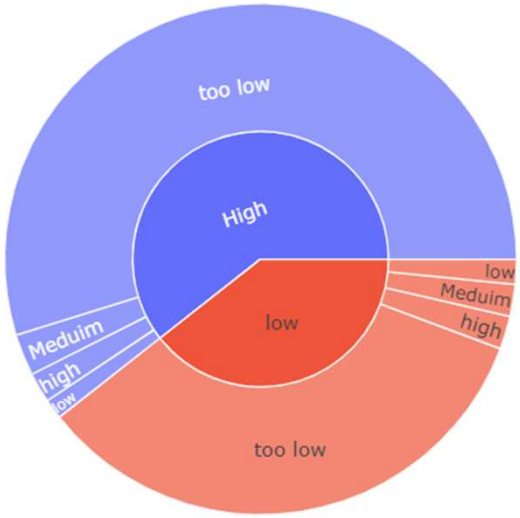


- 519 out of the high traffic recipes are too low sugar levels
- Only 18 has high sugar levels



- 319 out of the low traffic recipes are too low sugar levels
- Only 15 has high sugar levels

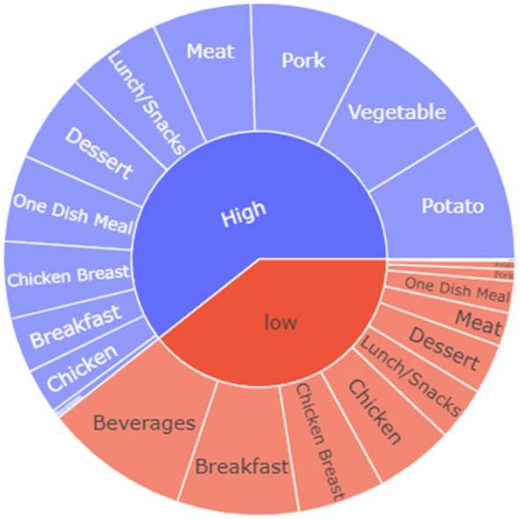
Traffic Vs Sugar categories



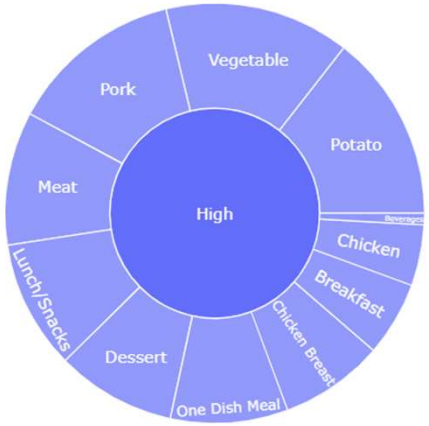
- High traffic 574
- Low traffic 373

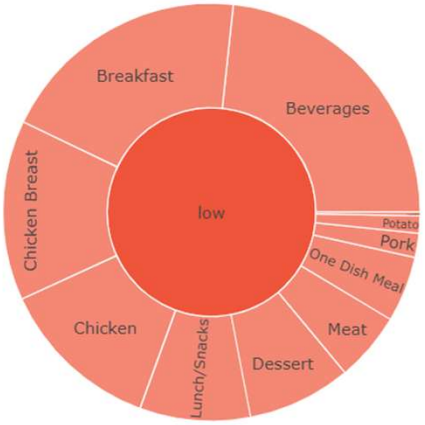
- We can notice that the sugar is doesn't defer to much in the high traffic or low traffic recipes
- We would further prove that with other tools
- That mean it's not a strong factor at all

Traffic Vs Category



- High traffic 574
- Low traffic 373

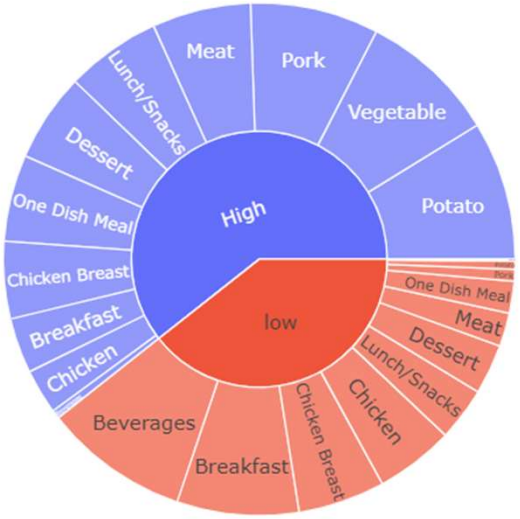




- 83 out of the high traffic recipes has the tag Potato
- Only 5 has tag Beverages(drinks)

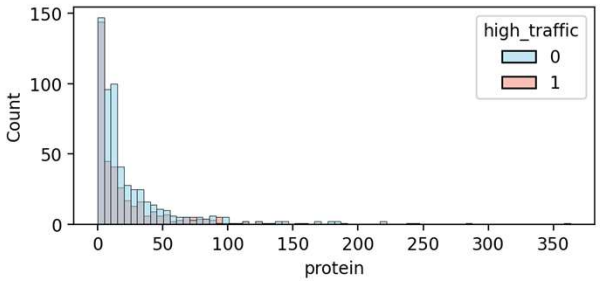
- Beverages with 87 recipe of the lowest traffic in the site
- Only 5 potato recipes are in the low rank

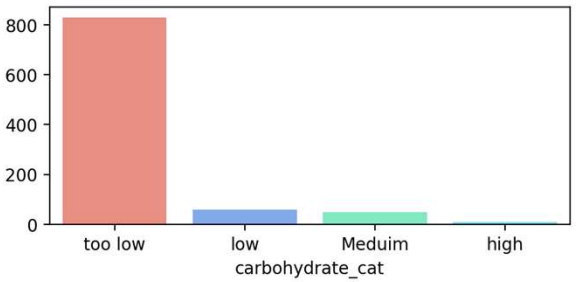
Traffic Vs Category

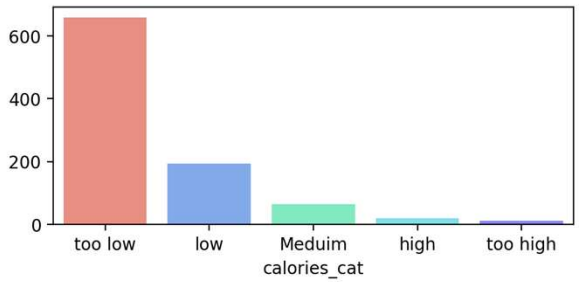


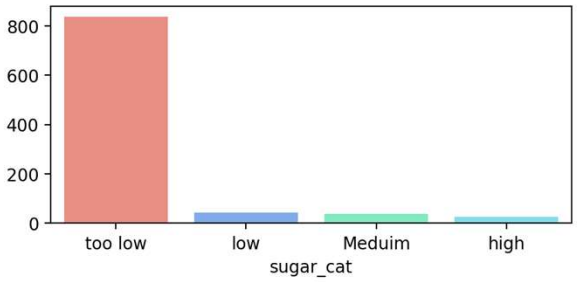
- High traffic 574
- Low traffic 373

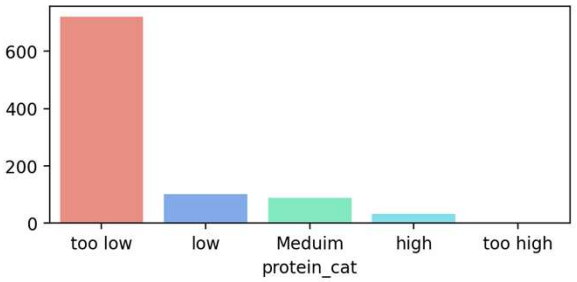
- We can notice that some categories of food is always high wile other categories is low most of the time
- We would further prove that with other tools
- That mean that the category feature can be so useful in determine the high traffic recipes

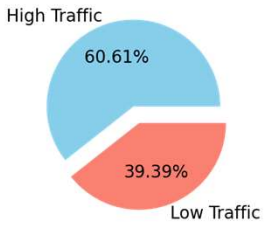

high_traffic
0
1
Count
protein

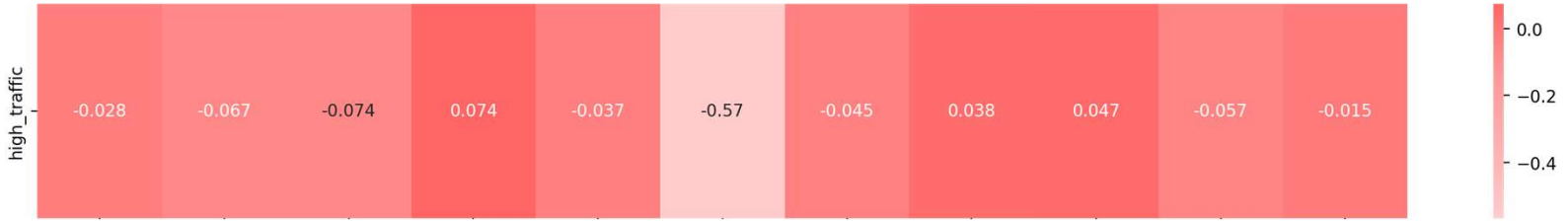

too low
low
Meduim
high
carbohydrate_cat


too low
low
Meduim
high
too high
calories_cat

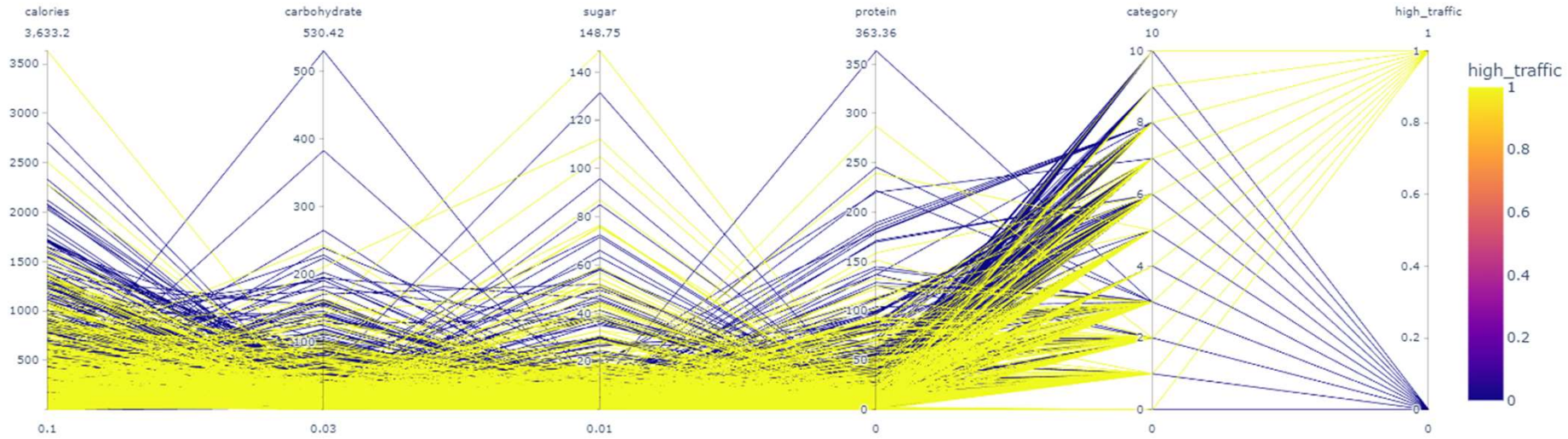

too low
low
Meduim
high
sugar_cat


too low
low
Meduim
high
too high
protein_cat

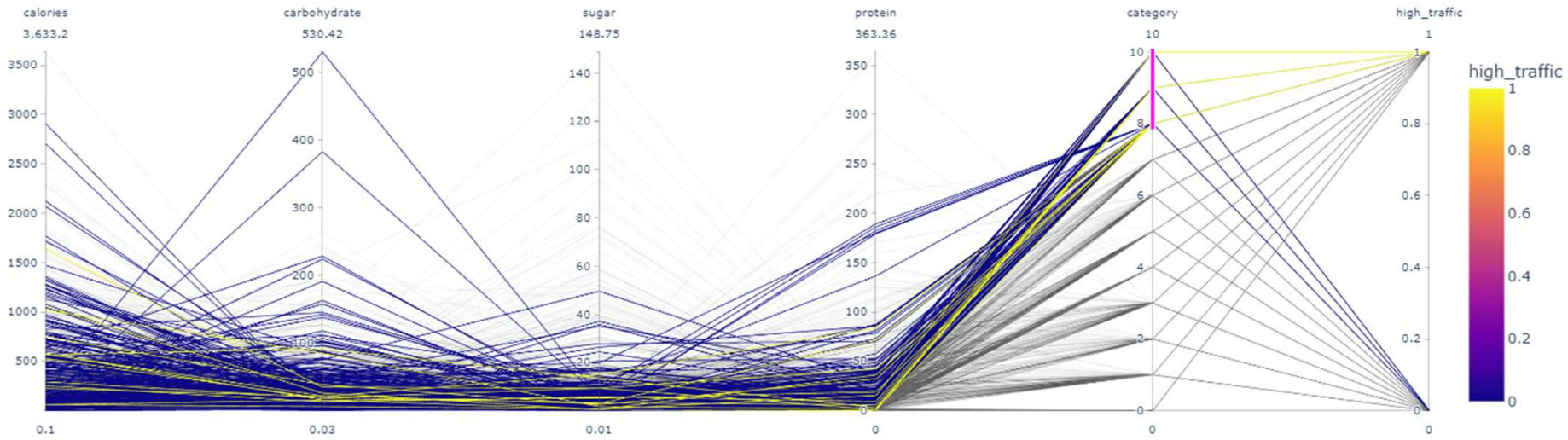

High Traffic
60.61%
39.39%
Low Traffic


high_traffic
-0.028
-0.067
-0.074
0.074
-0.037
-0.57
-0.045
0.038
0.047
-0.057
-0.015
0.0
−0.2
−0.4

Parallel coordinates plot(PCP)

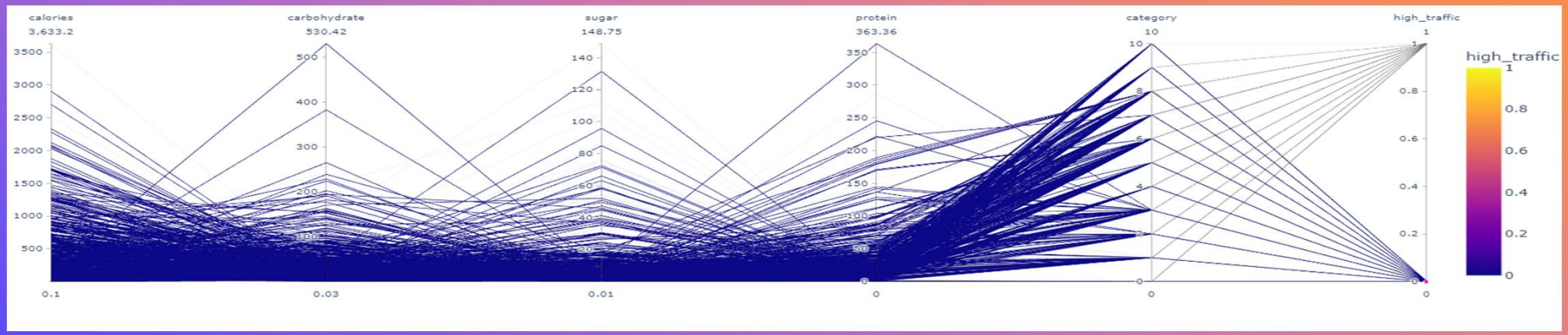


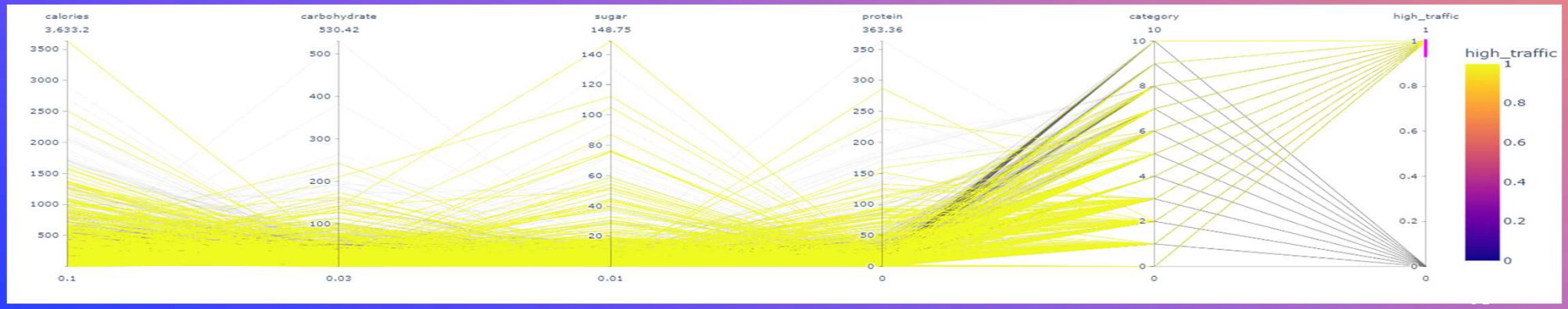
Parallel coordinates plot(PCP)



PARALLEL COORDINATES PLOT(PCP)

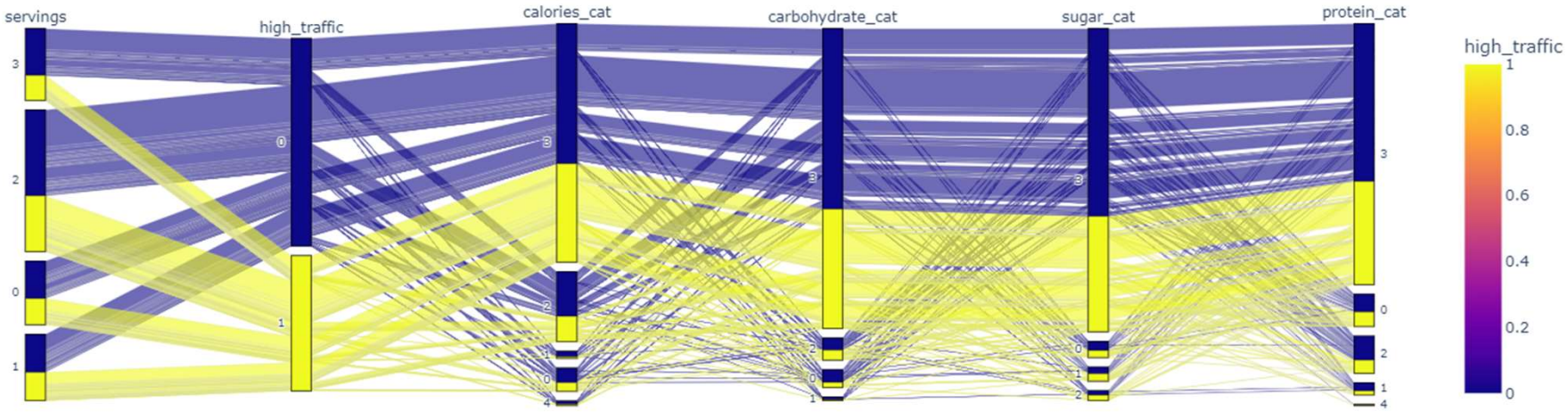
0: HIGH TRAFFIC (BLUE)





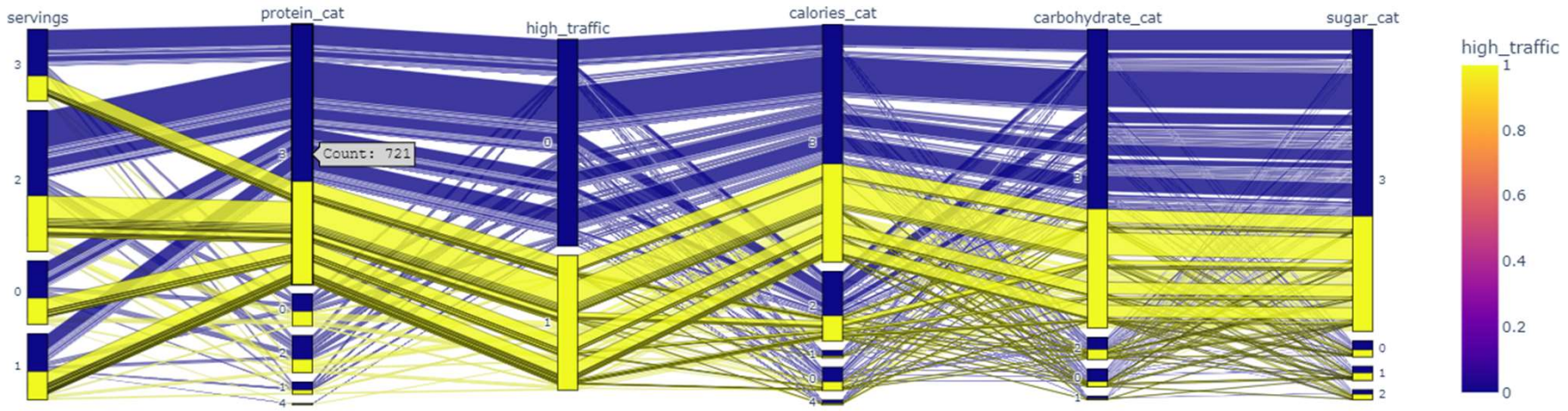
PARALLEL COORDINATES PLOT(PCP)

0: HIGH TRAFFIC (BLUE)



PARALLEL COORDINATES PLOT(PCP)

0: HIGH TRAFFIC (BLUE)



- We can see how protein can affect the high traffic for example for protein category 3 ‘too-low’ affect highly in how the traffic increase

MODELS AND BUSINESS METRICS

Business metric

We care more for the prediction of the recipes that can achieve high traffic more than the over all accuracy

Which will be measured as number of predicted as Hight traffic recipes and was a high traffic one over the total number of all the high traffics recipes that our model predicted

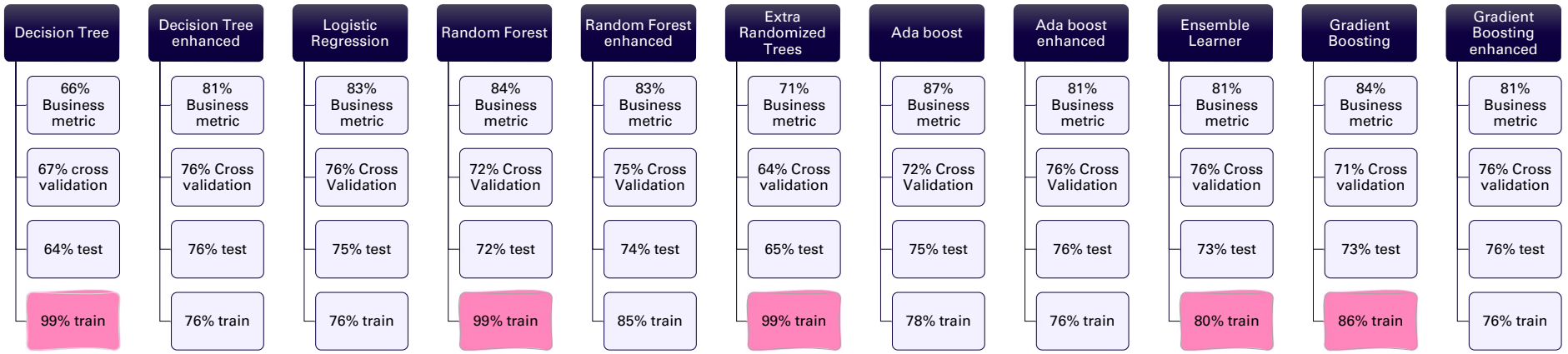
Out of all high traffic recipes how many were correct
This metric is the Recall

Business metric

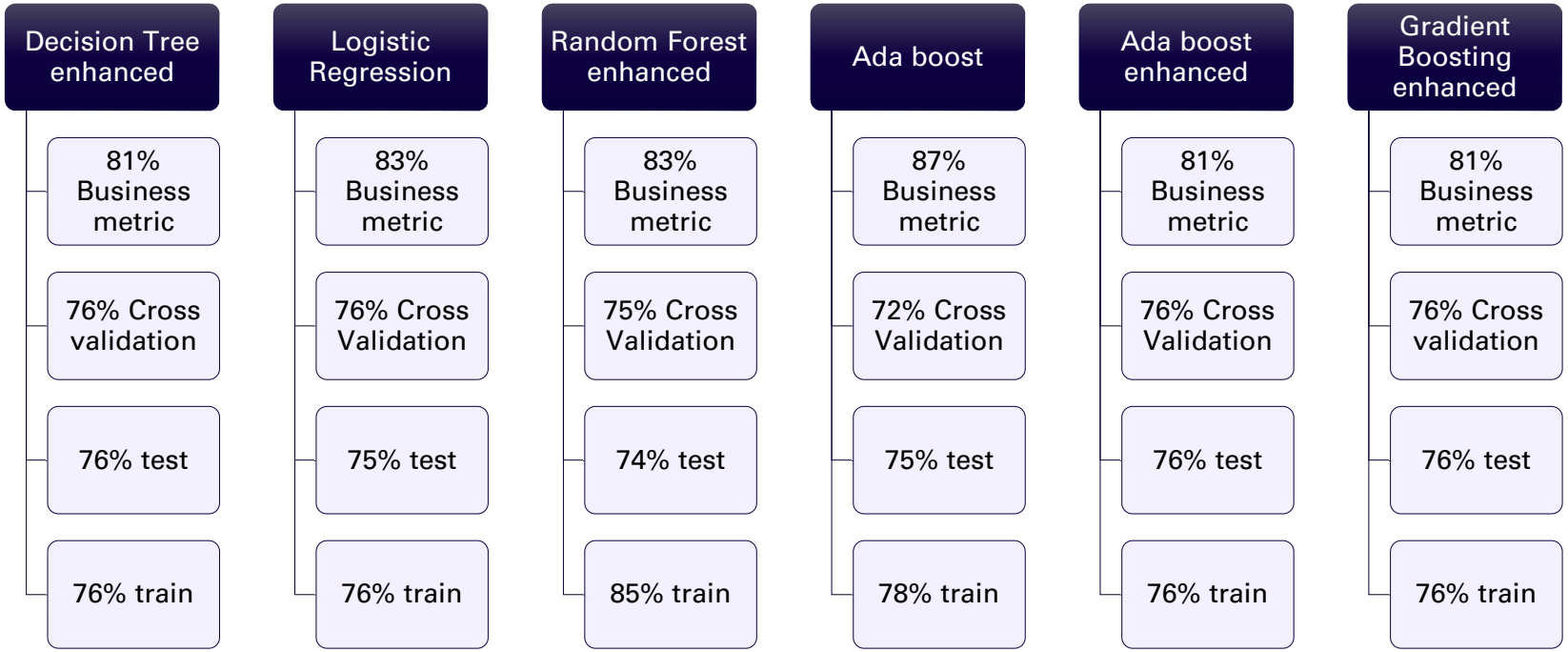
Out of all high traffic recipes how many were correct
This metric is the Recall

We need at least to reach 80% or higher to achieve the KPI and increase the traffic on the website

Models

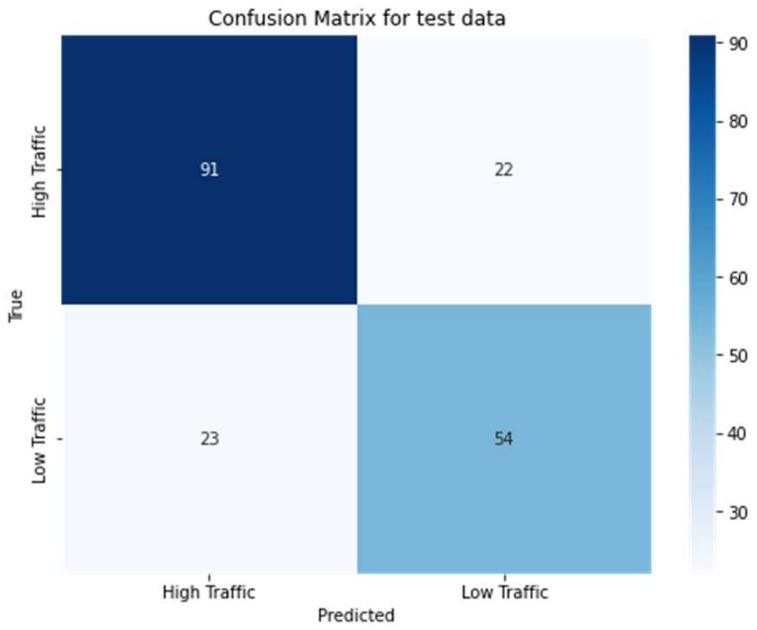

Decision Tree
66% Business metric
67% cross validation
64% test
99% train
Decision Tree enhanced
81% Business metric
76% Cross validation
76% test
76% train
Logistic Regression
83% Business metric
76% Cross Validation
75% test
76% train
Random Forest
84% Business metric
72% Cross Validation
72% test
99% train
Random Forest enhanced
83% Business metric
75% Cross Validation
74% test
85% train
Extra Randomized Trees
71% Business metric
64% Cross validation
65% test
99% train
Ada boost
87% Business metric
72% Cross Validation
75% test
78% train
Ada boost enhanced
81% Business metric
76% Cross Validation
76% test
76% train
Ensemble Learner
81% Business metric
76% Cross validation
73% test
80% train
Gradient Boosting
84% Business metric
71% Cross validation
73% test
86% train
Gradient Boosting enhanced
81% Business metric
76% Cross validation
76% test
76% train

Models


Decision Tree enhanced
81% Business metric
76% Cross validation
76% test
76% train
Logistic Regression
83% Business metric
76% Cross Validation
75% test
76% train
Random Forest enhanced
83% Business metric
75% Cross Validation
74% test
85% train
Ada boost
87% Business metric
72% Cross Validation
75% test
78% train
Ada boost enhanced
81% Business metric
76% Cross Validation
76% test
76% train
Gradient Boosting enhanced
81% Business metric
76% Cross validation
76% test
76% train

Models

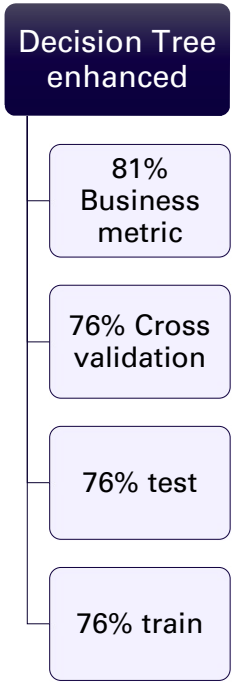
- Decision Tree enhanced
  - 81% Business metric
  - 76% Cross validation
  - 76% test
  - 76% train

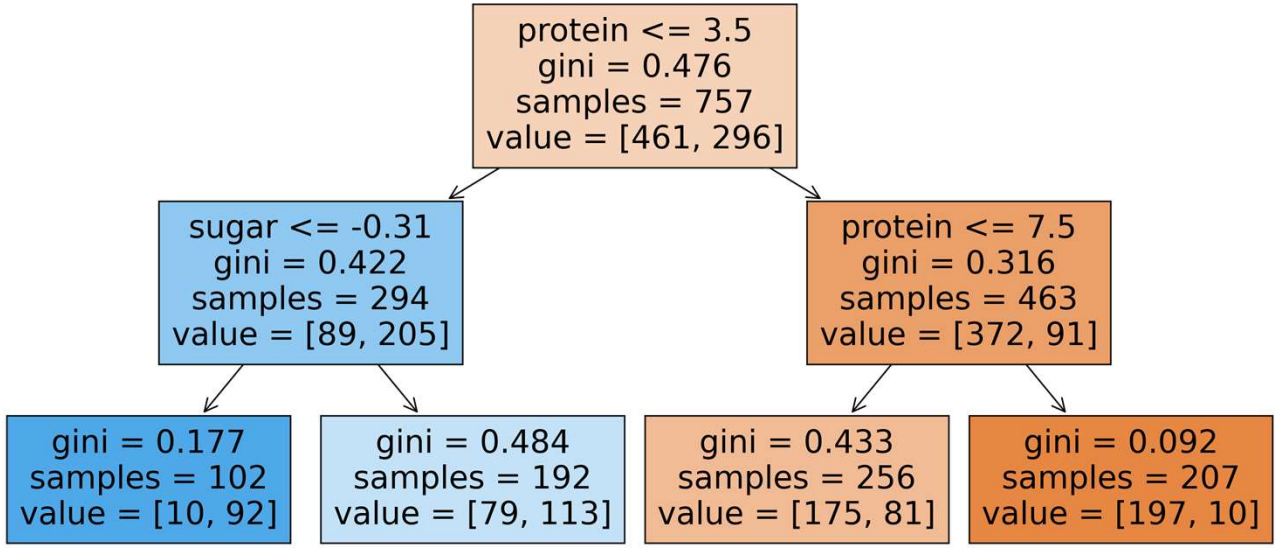


Business metric 81%
Recall of class high

Recall of class low traffic 70%

Models

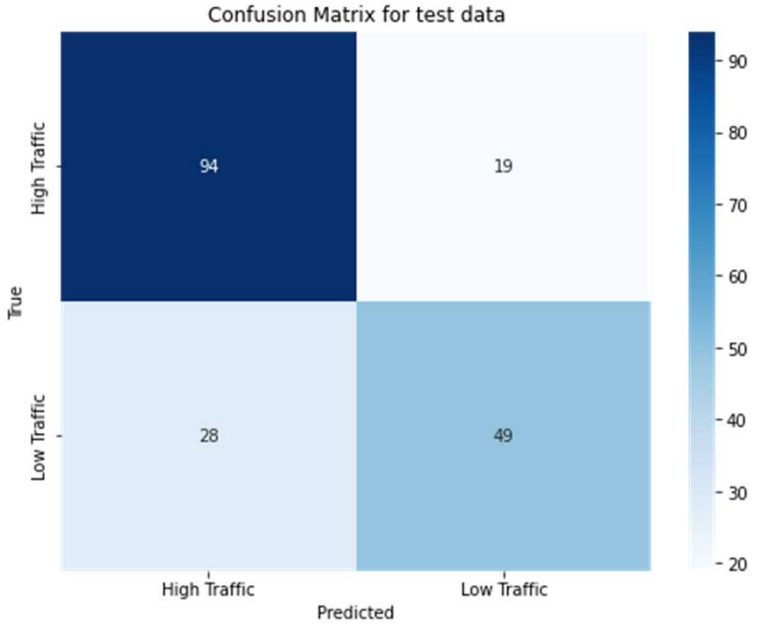

Decision Tree enhanced
81% Business metric
76% Cross validation
76% test
76% train


protein <= 3.5
gini = 0.476
samples = 757
value = [461, 296]
sugar <= -0.31
gini = 0.422
samples = 294
value = [89, 205]
protein <= 7.5
gini = 0.316
samples = 463
value = [372, 91]
gini = 0.177
samples = 102
value = [10, 92]
gini = 0.484
samples = 192
value = [79, 113]
gini = 0.433
samples = 256
value = [175, 81]
gini = 0.092
samples = 207
value = [197, 10]

Models

Logistic Regression

- 83% Business metric
- 76% Cross Validation
- 75% test
- 76% train

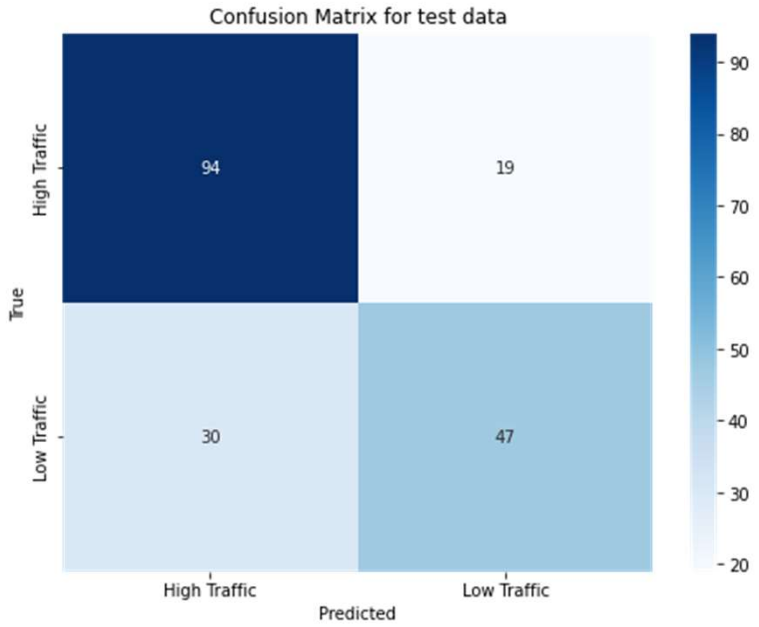


Business metric 83%
Recall of class high

Recall of class low traffic 64%

Models

- Random Forest enhanced
  - 83% Business metric
  - 75% Cross Validation
  - 74% test
  - 85% train

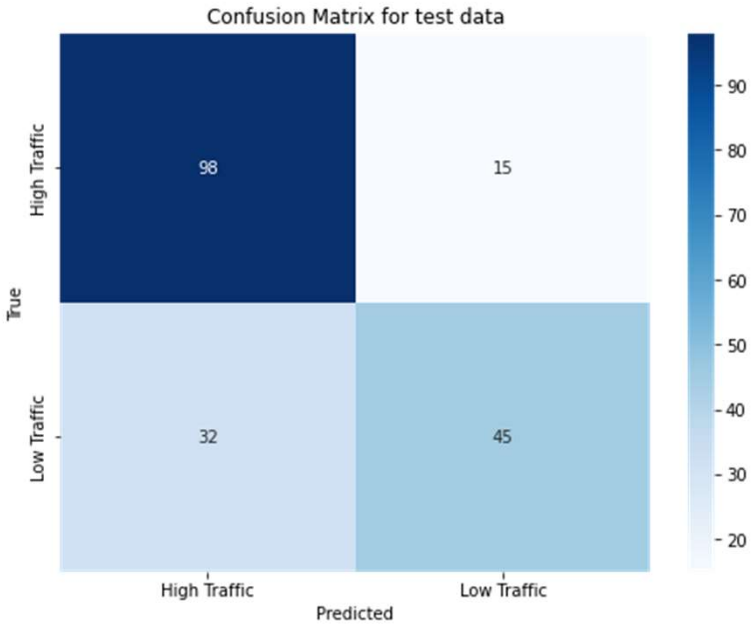


Business metric 83%
Recall of class high

Recall of class low traffic 61%

Models

- Ada boost
  - 87% Business metric
  - 72% Cross Validation
  - 75% test
  - 78% train

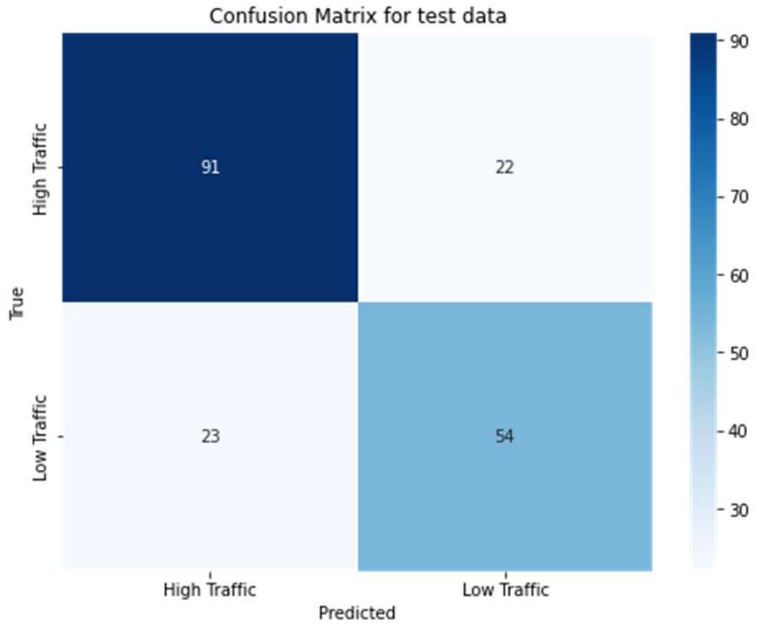


Business metric 87%
Recall of class high

Recall of class low traffic 58%

Models

- Ada boost enhanced
  - 81% Business metric
  - 76% Cross Validation
  - 76% test
  - 76% train

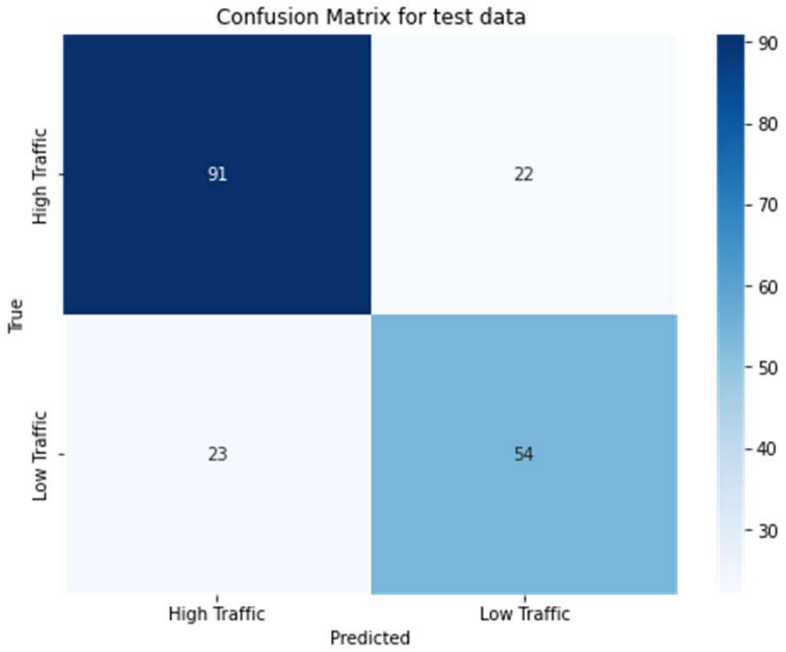


Business metric 81%
Recall of class high

Recall of class low traffic 70%

Models

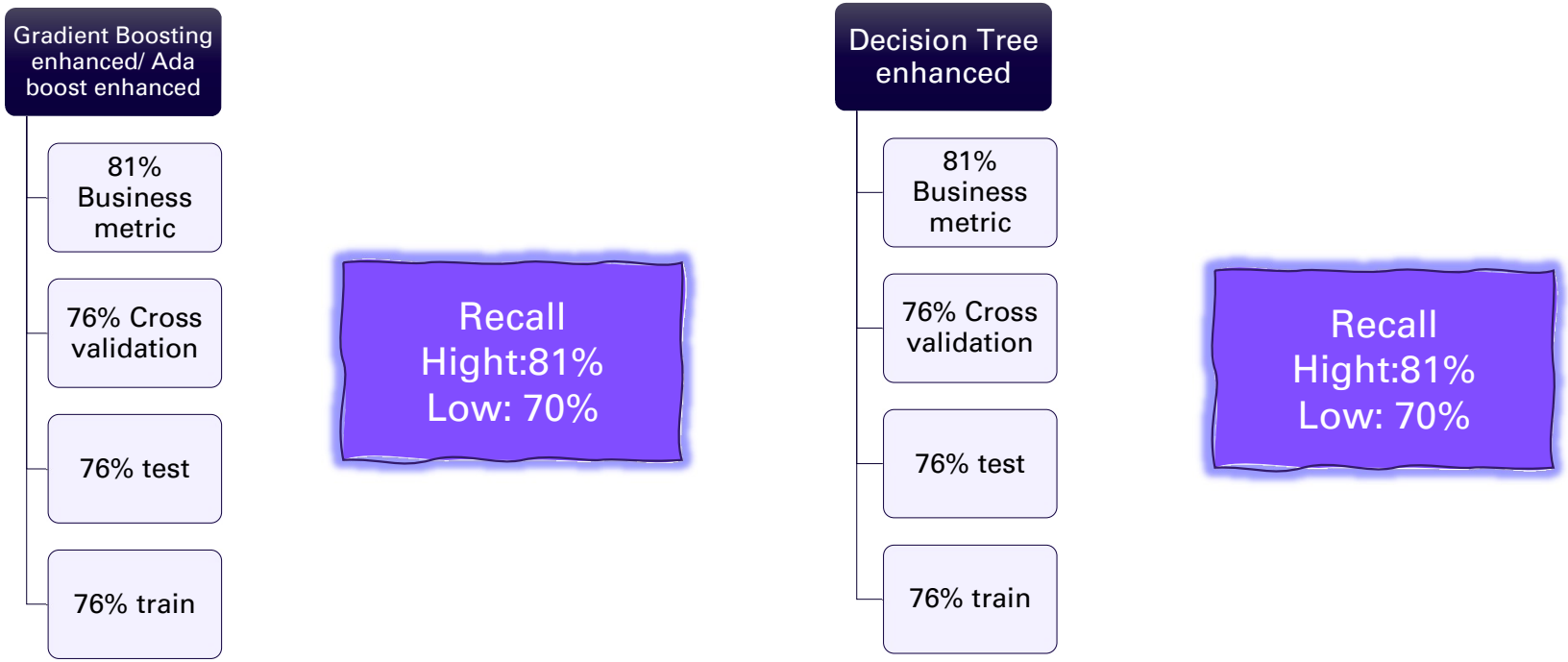
- Gradient Boosting enhanced
  - 81% Business metric
  - 76% Cross validation
  - 76% test
  - 76% train



Business metric 81%
Recall of class high

Recall of class low traffic 70%

Models *Finals*



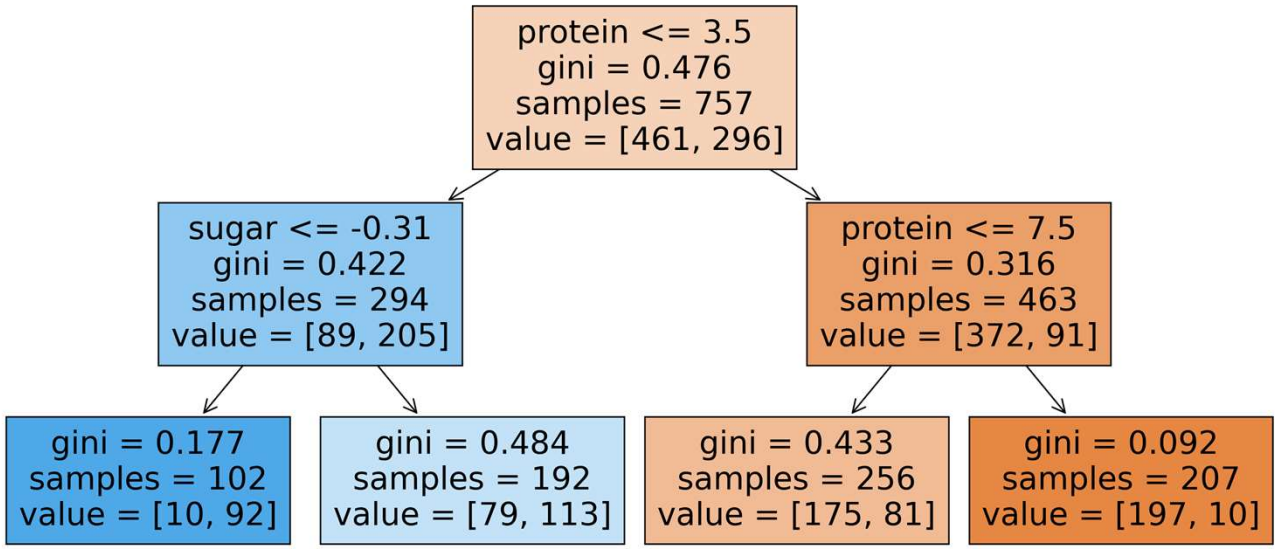
WHAT MODEL SATISFY THE KPI

Decision Tree enhanced

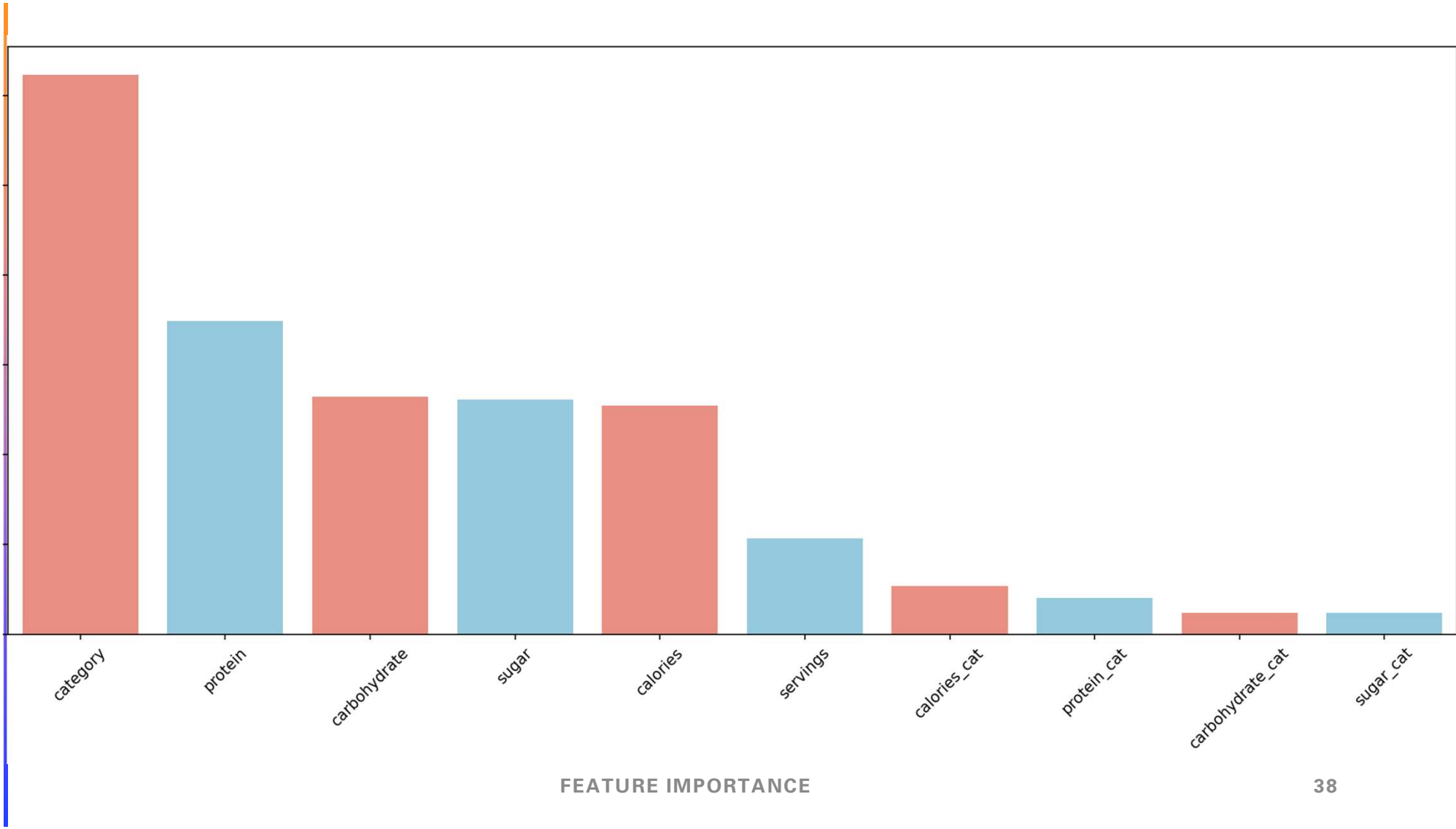
- 81% Business metric
- 76% Cross validation
- 76% test
- 76% train

Recall
Hight:81%
Low: 70%

Easy to interpret how the decision is made



FEATURE IMPORTANCE

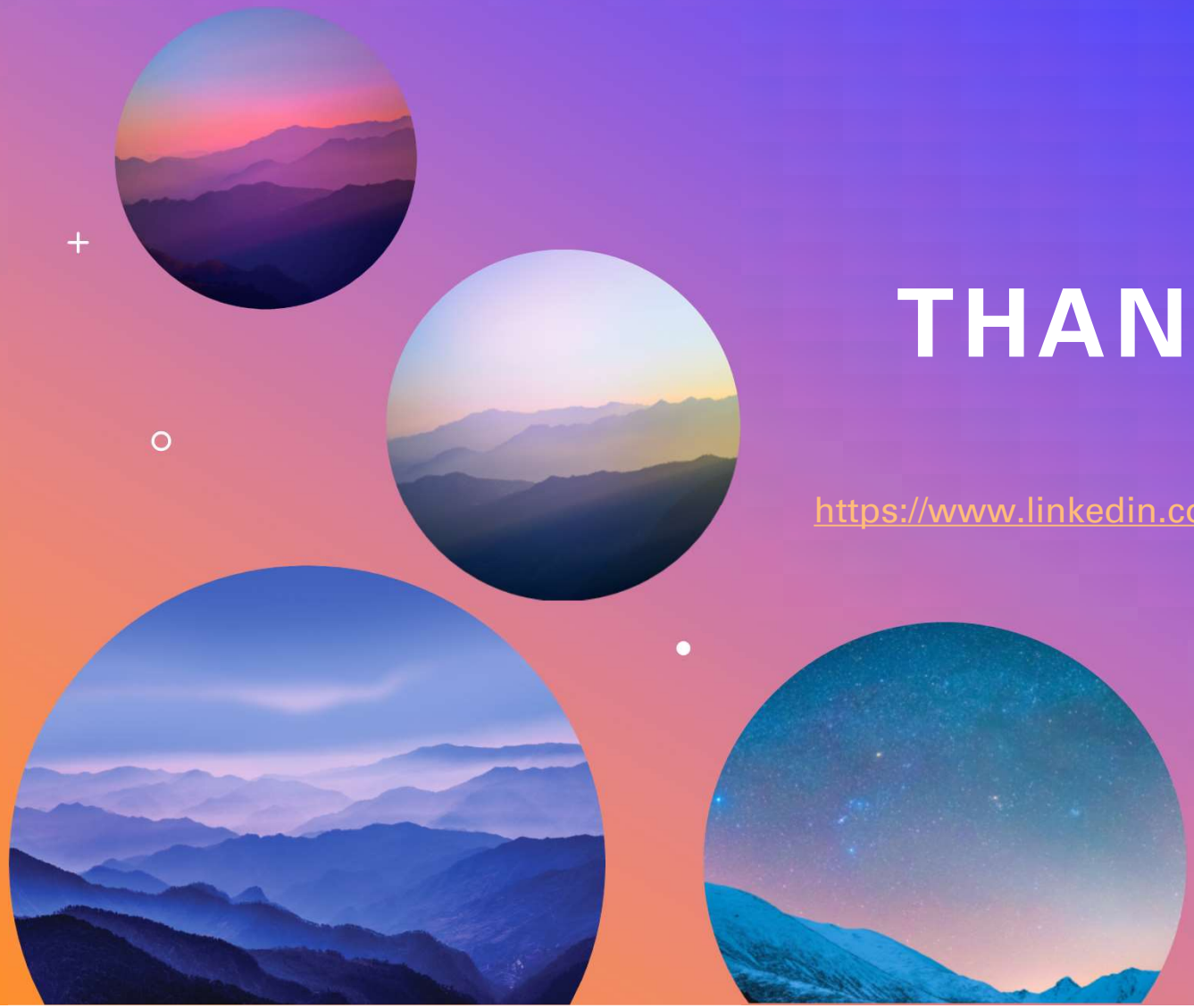


FEATURE IMPORTANCE

RECOMMENDATIONS

- We need to personalize the user experience ; Tasty Bytes aim is to lead the world and provide inspiration in the field of food recipes and delivery.
- But people taste defer and we need to make sure that every user experience is enjoyable and customized for that user.
- That will make the user engaged more and even he will recommend this experience to his friends .
- That will help in increasing the traffic and the time people spending in the website this is so important factor.
- What I recommend ?
  - Building a Recommendation system in your website which will make each user experience a unique experience of what he really want to see.
  - Additionally, I suggest focusing on the features that greatly influence people's preferences. As we discussed earlier, some categories are more popular than others.

HIGHT TRAFFIC



THANK YOU

Hossam Ahmed Salah

https://www.linkedin.com/in/hossamahmedsalah/