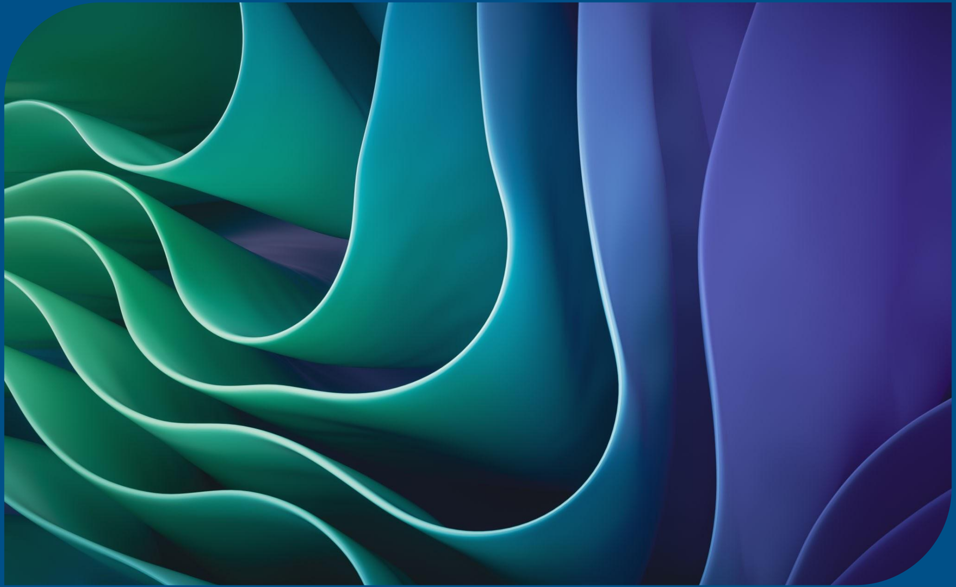
12.01.24

High Traffic Recipes

Findings and Suggestions



Tucker Rhodes

Data Validation and Cleaning

DOCUMENTATION

- 8 columns with 947 observations

VALIDATION

- Missing Values
  - 52 missing values in calories, carbohydrate, sugar, and protein
  - 373 missing values in high_traffic
- Duplicate rows
  - No duplicate values
- Data Types
  - Category, servings, and high_traffic were changed from object to category

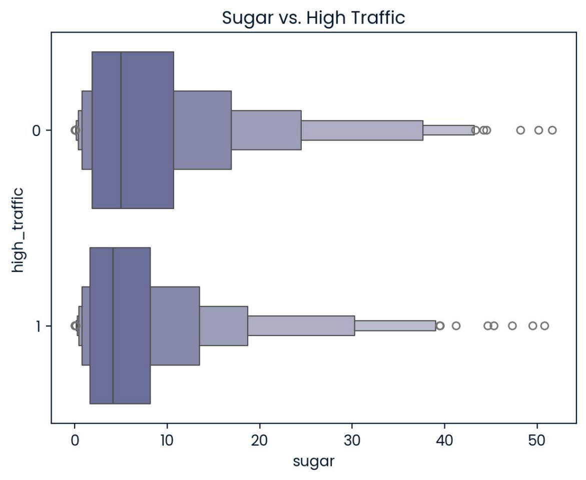
CLEANING

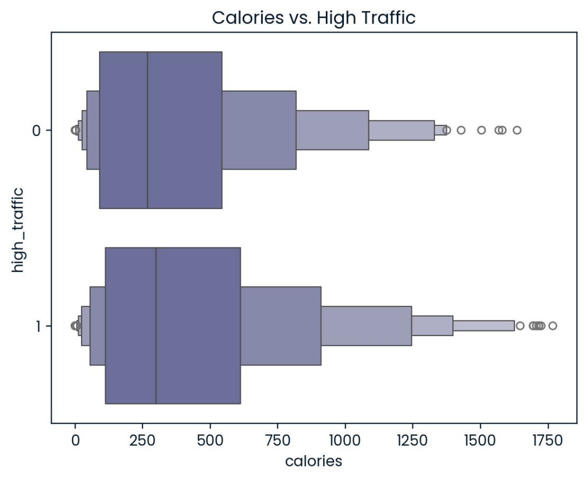
- Dropped recipe column
- Dropped outliers
  - Outliers were defined by being more than 3 standard deviations from the mean of the column
- Empty values for high_traffic were filled with “Not High”
  - After this rows containing any missing values were dropped
- The classes “High” and “Not High” were changed to “1” and “0” respectively

Data Analysis: Decimal Columns

Goal: For each of the decimal columns, is there evidence that they influence traffic?

- Columns containing decimal values, namely the calories, carbohydrate, sugar, and protein columns, were analyzed using box and whisker plots and subjected to statistical testing
- Statistical testing showed that there was evidence that sugar had a statistically significant impact on traffic
- As for the others, we simply cannot say one way or the other

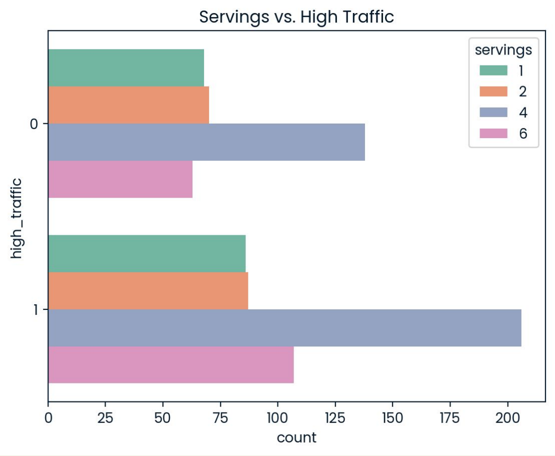


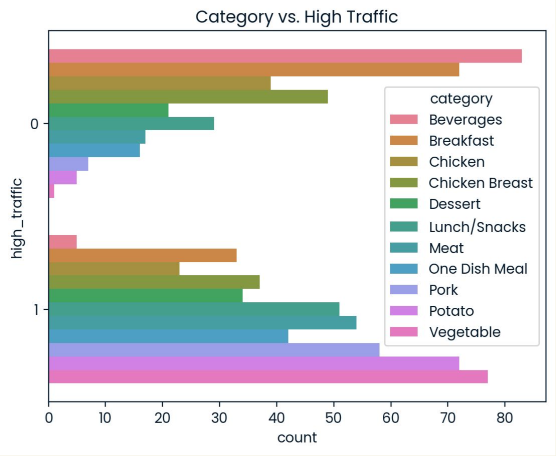


Data Analysis: Categorical Columns

Goal: For each of the categorical columns, is there evidence that they influence traffic?

- The servings and category columns were analyzed with a bar chart that where the colors represented the servings size or meal category
- "Shape" = shape the bars make for each class
- Shape was similar between classes for servings - likely not a significant driver
- Shape was significantly different for the category column
  - When the recipe category is "Beverage", it almost always fails to bring high traffic
  - When the recipe category is "Vegetable", it almost always brings high traffic
- Category appears to have a significant impact on traffic while servings likely only has a small impact if any at all





Model Development: Metrics and Baseline

Statement of Problem

- The business wants to predict that a recipe will bring high site traffic 80% of the time.
- In machine learning, this type of problem is called a binary classification problem because we are attempting to put observations into one of two classes.

Metrics Used to Evaluate

- Accuracy: What percentage of the predictions were correct?
- Precision: What percentage of High Traffic predictions were actually High Traffic recipes?
    - How confident can we be when the model says a recipe will bring high traffic?
- Recall: What percentage of the actual High Traffic recipes did the model predict correctly?
    - When a recipe will bring high traffic, how confident can we be that the model will predict it as such?

Baseline Model Evaluation

- Accuracy: 76%
- Precision: 78%
- Recall: 79%

Model Development: Comparison Model

Model/Data Tuning

- Tuning baseline model parameters, and transforming data did not provide metrics surpassing those of the baseline
- Best path was to train a different type of model

Model 2

- Model 2 used a different algorithm and works best with scaled data
- After scaling the data and tuning the model parameters the best model had the following metrics:
    - Accuracy: 76%
    - Precision: 79%
    - Recall: 78%
- Changes:
    - Accuracy: +/-0%
    - Precision: +1%
    - Recall: -1%

Business Metrics: What to track and how to track it

What metric should the business track?

- The business needs to track how well the model is able to predict high traffic recipes
- Precision would be the easiest to track and will give us insight into how well the model predicts high traffic recipes

How can the business track this?

1. If there is not an explicit number to define a recipe as "high traffic", this should be determined first
2. Determine a time interval to check if a posted recipe has brought high traffic (at least a week recommended)
3. Post recipe recommendations of the model to the website
4. When it is time to switch recipes, record which ones brought high traffic and which ones did not
5. At the end of the chosen time interval, take the number of recipes that had high traffic and divide that number by the number of recipes posted to the home page. This will give you the percentage you want to track
6. Repeat steps 3-5 for the next iteration of recipes

Note:

If the precision ever drops below 60%, notify the data team at your earliest convenience. Certainly feel free to notify us sooner in the event precision begins to drop but at this point the model is only slightly better than guessing and retraining with new data may be needed.

Model Improvement

Improvement Opportunities

- Rerun data with all values in ‘high_traffic’ column
- Gather more data
- Collect a more balanced data set

Final Recommendations

Model Recommendation

- Use model 2 and evaluate its effectiveness
    - More confidence (precision)
    - Is likely to correctly predict 4 out of every 5 high traffic recipes

Practice Recommendations

- Implement the procedure outlined on the Business Metrics slide as a part of the recipe rotation procedure
- Continue to collect data
    - Evaluate model
    - Improve model
- Implement recipes that with one of the top 3 categories
    - Vegetables
    - Potato
    - Pork (though this may fall in the generic meat category which was ranked 4th)
- Less sugar appears to be correlated with higher traffic