Estimating Calories and Nutritional Inter-dependencies in Starbucks' menu

Khevna Parikh
kp2936@nyu.edu

Dhruv Saxena
ds6802@nyu.edu

1 Motivation

Starbucks, the world's largest coffeehouse chain, has had an impressive growth trajectory since its founding in 1971 with Seattle's historic Pike Place Market(**?**). The store quickly became popular for its coffee beans, tea, and spices, and has expanded globally to an impressive 36,160 stores in over 80 countries, 45 percent of which are located in the United States. With a market share of approximately 37 percent and revenue of $32 billion, Starbucks' sales and revenue growth remain unmatched. The company's portfolio is projected to grow 7% to 9% in the upcoming fiscal years, demonstrating the company's commitment to continued expansion and success.

Starbucks is known for serving a wide variety of hot and cold drinks, ranging from coffee and espresso beverages to teas and refreshers. In addition, customers can create their own personalized drinks by selecting from a variety of options such as adding extra espresso shots, flavored syrups, or different types of milk. Apart from beverages, the company also offers whole-bean coffee, pastries, and other savory food items. With such a diverse selection, Starbucks caters to a wide range of tastes and preferences, making it a popular choice among consumers around the world.

Given its sizeable presence, we aim to explore the nutritional value of Starbucks' menu items. Specifically, we will examine the **calorific value** of its food and drink items and explore any **inter-dependencies with nutritional components** like sugar and cholesterol levels.

2 Dataset

In our report, we are exploring three different datasets related to Starbucks: the Starbucks drink menu, the Starbucks food menu, and the Starbucks expanded drink menu (Link to Dataset). Each dataset provides nutritional information, such as calories, fat, carbohydrates, fiber, and protein. The expanded drink menu goes further to include information on sugar, trans fat, saturated fat, iron, caffeine, cholesterol, vitamins, and sodium levels for each drink item.

The expanded drink menu dataset contains information for 242 drinks and includes nutritional facts based on the size and type of milk used. It covers four different kinds of milk - soy milk, nonfat milk, 2% milk, and whole milk - and nine different beverage categories, as shown in Section **??**. It also covers Starbucks' two sizes of espresso shots, "Solo" and "Doppio," as well as its regular beverage sizes of "Short," "Tall," "Grande," and "Venti." It's important to note that Starbucks offers more drink varieties and customizations beyond what is captured in this dataset. This dataset is well-curated for our problem statement, despite having some missing values.

On the other hand, Starbucks' drink menu is limited and to some degree, messy. It only provides nutritional information for a 12-ounce or their "Tall" drink, and half the data is missing, marked with a dash, "- ". Instead of eliminating these items, we will compare the missing beverages to that of the expanded drink menu to retain as much information as possible. However, in the case where nutritional information is listed, many beverages have a fat, fiber, or protein level of 0.

The last dataset contains nutritional information for Starbucks food menu items. The small-scaled dataset includes calories, fat, carbohydrates, fiber, and protein for 133 food items, with no missing values.

3 Methodology

In this study, we aimed to examine the relationship between nutritional components and calorie content among menu items at Starbucks. To achieve this, we utilized both hypothesis testing and regression analysis. Hypothesis testing allowed us to determine the correlation between the independent variables (nutritional components) and the outcome variable (calorie content), and establish statistical significance amongst these variables. On the other hand, regression analysis allowed us to develop a predictive model that could estimate calorie content based on the nutritional components of Starbucks drinks. These findings were validated with

the findings from hypothesis testing to provide a comprehensive overview of the nutritional value of Starbucks menu items.

3.1 Hypothesis Testing Framework (Part 1)

The correlation coefficient, r, measures the strength and direction of the linear relationship between two variables. To determine if the linear relationship is strong enough between the predictor and outcome variables, we use a hypothesis test of the significance of the correlation coefficient.

We started by establishing our null hypothesis, which assumed there is no correlation between the nutritional features and calorie content, and an alternative hypothesis that assumed there is a correlation.

A t-test is used to determine whether there is a significant difference between the means of two groups. In the context of correlation analysis, we used the t-statistic to measure the significance of the correlation between these variables. (Kumar, 2022) We calculated the test statistic using the formula $t = (r * \sqrt{n-2})/\sqrt{1-r^2}$, where r represents the sample correlation coefficient and n represents the sample size. We chose a commonly used value of alpha = 0.05 for the significance level (α). This indicates a 5% chance of rejecting the null hypothesis when it is true.

We note that correlation does not necessarily imply causation or independence, but it measures the existence of a linear relationship between two variables. If two features are independent, then their correlation will be zero. Nonetheless, by conducting this hypothesis test, we aimed to identify any significant correlations between nutritional features and calorie content of Starbucks menu items. In the following section, we will describe the data preprocessing steps taken after gathering the data.(Starbucks)

3.2 Data Preprocessing

To ensure that our dataset was of high quality and utility, we performed a variety of data cleaning and processing steps. First, we renamed columns to make them easier to work with, such as converting "Trans Fat (g)" to "trans_fat".

We then dealt with data typos, such as for "Total Fat" which had a row value of "3 2" instead of "3.2". We also removed the percentage sign from certain columns like Iron or Vitamin A, which had it in the row value. In addition, Tazo Tea Drinks had caffeine values of "varies" or "Varies", which we replaced with "NaN" and removed these rows during modeling due to the lack of information on these drinks. Each drink can be customized and determining the exact calorie amount for these drinks was difficult to obtain.

Next, we changed the data types of Total Fat, Vitamin A, Vitamin C, Calcium, Iron, and Caffeine columns from object to float to make them numerical data types. We also converted Sodium, Cholesterol, and Caffeine columns from milligrams to grams to make them more easily comparable with the rest of the nutritional columns, which were in grams.

To prepare our dataset for predictive modeling, we addressed outliers using a z-score threshold method and removed any outliers that were outside of 3 standard deviations. We then moved on to address the issue of multicollinearity, which can arise when two or more independent variables are highly correlated, making it difficult to determine which variables are most important in predicting the calorific content of Starbucks drinks.

To detect multicollinearity, we calculated the variance inflation factor (VIF), which measures the extent to which the variance of a predictor variable is increased due to multicollinearity with other predictor variables in the model. A high VIF indicates a large standard error for the parameter estimate associated with the variable and is a necessary but not sufficient condition for evidence of pairwise correlation (Bohn and Stein, 2009). We found that the sugar and cholesterol variables were highly correlated, which could lead to multicollinearity issues in our models. Sugar is know to be a direct source of calories, and since our objective to estimate the calorific value of Starbucks drinks, it was more relevant to keep the sugar variable and drop the cholesterol feature.

3.3 Hypothesis Testing Framework (Part 2)

To conduct hypothesis testing, we first removed the 'Beverage_category', 'Beverage', and 'Beverage_prep' variables from our dataset, as our main goal was to investigate the inter-dependencies between the nutritional components and calories. We started by computing the Pearson's correlation coefficient between calories and each of the nutritional component features.

Using the formula $t = (r * \sqrt{n-2})/\sqrt{1-r^2}$, we computed the test statistic or the t-value for each pair of variables. We calculated the two-sided

p-value for the t-distribution with 240 degrees of freedom using the resulting test statistic.

After finding the corresponding p-value, we examined whether or not the correlation between these two variables was statistically significant. We will discuss the results of the hypothesis testing in the following section.

3.4 Predictive Modeling

To predict the calorie content among menu items at Starbucks, we followed several more processing steps. Firstly, we split the data into training and testing sets to avoid data leakage. Specifically, we split the data into a 75:25 ratio, where 75% of the data was used for training and 25% for testing.

The dataset contains three categorical columns: 'Beverage_category', 'Beverage', and 'Beverage_prep'. There are nine unique values for 'Beverage_category', 33 unique values for 'Beverage', and 13 unique values for 'Beverage_prep'. One-hot encoding these features would result in a large number of columns. To address this problem, we used the hashencoder technique, which involves applying a hash function to each category and mapping it to a fixed-size value.(Weinberger et al., 2009) This technique reduces the dimensionality of the feature space compared to one-hot encoding, allowing us to avoid the problem of high cardinality.

3.4.1 Linear Regression

In our feature selection process, we followed Jason Brownlee's recommendation and used the f_regression function from the scikit-learn library, along with the SelectKBest method (Brownlee, 2020). To determine the most relevant features for modeling, this method calculates the cross-correlation between each regressor and the target variable using the r_regression method. Then, this correlation is transformed into an F score and a p-value. These values are used to rank the features based on their correlation with the target variable and select the top top k features that showed the highest correlation scores. (Learn) We opted to use the SelectKBest method, which assigns scores to each feature based on the scoring function, and then selects the k features with the highest scores.

We used the Mean Absolute Error (MAE) to evaluate the performance of our regression model. The MAE measures the average absolute difference between the predicted and true values of the target variable, and it is less sensitive to outliers. Lastly, since our dataset is relatively small, we use RepeatedKFold to obtain a better representation of the data. This is a cross-validation method similar to the K-Fold cross-validation method, but it repeats the process multiple times with different random splits of the data.

3.4.2 Gradient Boosting on Decision Trees

In addition to linear regression, we employed an XGBoost model to predict the calorie content among menu items at Starbucks. We followed a similar approach to the linear regression method by splitting the data into training and testing sets in a 75:25 ratio and encoding the three categorical variables using hashencoder.

We performed hyperparameter tuning by optimizing the following parameters: 'max_depth' (maximum tree depth), 'learning_rate', 'n_estimators' (number of gradient boosted trees), and 'colsample_bytree' - which defines what percentage of features (columns) will be used for building each tree. However, the XGBoost algorithm can only handle numerical features, which was a limitation for our dataset. To overcome this, we employed CatBoost, another gradient boosting algorithm that can handle categorical features.

To optimize performance of CatBoost,, we conducted hyperparameter tuning on the following parameters: 'iterations' (maximum number of trees), 'learning rate', 'depth' (depth of the tree), and 'l2_leaf_reg' (coefficient at the L2 regularization term of the cost function). The parameters we tested for both XGBoost and CatBoost models are listed in the tables below.

XGBoost Parameters

Parameter	Values
max_depth	3, 6, 10
learning_rate	0.01, 0.05, 0.1
n_estimators	100, 500, 1000
colsample_bytree	0.3, 0.7

Catboost Parameters

Parameter	Values
iterations	100, 150, 200
learning_rate	0.03, 0.05, 0.1
depth	2, 4, 6, 8
l2_leaf_reg	0.2, 0.5, 1, 3

, and we employed grid search cross-validation to identify the optimal hyperparameters for the models. Finally, we selected the model that yielded

the lowest Root Mean Squared Error (RMSE) and fitted it to the test data.

4 Results

Figure 1: Results for the Hypothesis Test of the significance of the Correlation Coefficient

	Level_0	Level_1	Correlation Coefficient	T-statistic	P-value
1	cholesterol	calories	0.937914	41.889299	0.00000
3	sugars	calories	0.908817	33.747588	0.00000
6	total_carbohydrates	calories	0.783966	19.563609	0.00000
13	calories	trans_fat	0.635440	12.749070	0.00000
15	protein	calories	0.546193	10.101455	0.00000
19	calories	calcium	0.485662	8.607067	0.00000
25	iron	calories	0.412398	7.012968	0.00000
28	calories	sodium	0.373547	6.238577	0.00000
29	vitamin_A	calories	0.370343	6.176514	0.00000
30	dietary_fiber	calories	0.369126	6.153006	0.00000
38	calories	saturated_fat	0.325883	5.340080	0.00000
54	vitamin_C	calories	0.213281	3.381958	0.00084
81	calories	caffeine	-0.046857	-0.726705	0.46811

As we see in Figure 1, all nutritional components expect for caffeine had a positive correlation with calorie content. Stronger correlations were observed for some pairs, such as cholesterol and calories ($\rho = 0.93$), while weaker correlations were observed for others, such as Vitamin C and calories ($\rho = 0.21$). Our t-statistic was large for all pairs expect for caffeine and calories, providing evidence against the null hypothesis. The p-value was less than 0.05 for all pairs expect for caffeine and calories.

In our study, we utilized three different predictive models to examine the relationship between nutritional components and calorie content among menu items at Starbucks. Firstly, the SelectKBest method for linear regression model resulted in 19 important features. The model was fitted using the training data on these 19 features, and the Mean Absolute Error (MAE) was approximately 5.9 for the test set.

Next, we employed XGBoost, a widely used and efficient model known for its high-performance capabilities. Using grid search cross-validation, we identified the best hyperparameters for our model as 'colsample_bytree': 0.7, 'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 1000. The model with the lowest root mean square error (RMSE) of approximately 16.9 was fitted on the test set, and achieved a MAE of 8.6.

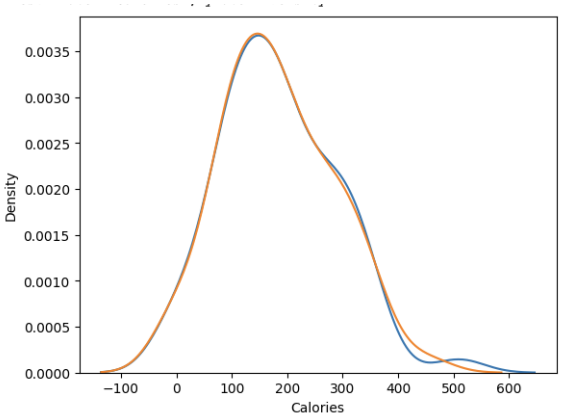
Lastly, we utilized CatBoost, a gradient boosting method for categorical features. The best hyperparameters for this model were identified as 'l2_leaf_reg': 3, 'learning_rate': 0.100, 'iterations': 200, and 'depth': 2. The MAE for this model was 9.1453, which was slightly higher than the XGBoost and linear regression model.

5 Analysis

Based on the results shown in Figure 1, we can observe that all nutritional components, except for caffeine, had a p-value of less than 0.05. Therefore, we can reject the null hypothesis for these pairs and find supporting evidence for the alternative hypothesis that there is a significant correlation between these nutritional components and calorie content of menu items at Starbucks.

Overall, the linear regression model provided the best accuracy in predicting calorie content among menu items at Starbucks. A model with a MAE of 5.9 indicates can be deemed as good. However, it is important to note that the the linear regression model fails to accurately predict certain types of drinks.

Figure 2: This figure depicts the density plots of the predictions generated by the linear regression model (represented by the blue line) and the actual calorie values (represented by the orange line). The overlap and alignment of the two lines indicate the accuracy of the model in predicting calorie content of menu items at Starbucks.



It was observed that the model struggled to accurately identify calorie content for soymilk-based drinks, caffe lattes, and caffe mochas. Furthermore, the high calorie drink White Chocolate Mocha was identified as a problematic menu item as depicted in Figure 3, as all models failed to predict its calorie content with the highest error.

This suggests that the nutritional components used in the model may not capture the complexities of these drinks to the full extent, and additional

factors may need to be considered.

6 Plan for Additional Analysis

First, we will examine the beverage prep column to identify why our models failed to detect soymilk-based drinks. It would also be valuable to gather data on portion sizes for each drink, as this can significantly impact the overall calorie content. We will also consider adding other Starbucks menu items, such as food, to our analysis.

To ensure the validity of our models, we will check for bias and confounding variables and explore different feature processing techniques like upsampling and downsampling. Furthermore, we are interested in analyzing the effects of reintroducing the cholesterol feature and removing sugars from our predictive modeling.

Next, we will use hypothesis testing to identify the most informative features for predicting calorie content. Since the correlation between caffeine and calorie content was found to be insignificant, it is worth exploring the implications of removing caffeine from our predictive modeling. We will also double-check our hypothesis test assumptions, such as linearity and normality amongst the nutritional components and calorie content. This could ensure that our predictive models results are reliable.

Lastly, we will consider the business implications of our findings and develop a marketing strategy for Starbucks based on the results of our analysis.

References

Jeffrey R. Bohn and Roger M. Stein. 2009. *Active credit portfolio management in practice*. Wiley.

Jason Brownlee. 2020. How to perform feature selection for regression data.

Ajitesh Kumar. 2022. Linear regression hypothesis testing: Concepts, examples.

Scikit Learn. Sklearn.feature_selection.f_regression.

Starbucks. Nutrition facts for starbucks menu.

Kilian Weinberger, Anirban Dasgupta, John Langford, Alex Smola, and Josh Attenberg. 2009. Feature hashing for large scale multitask learning. *Proceedings of the 26th Annual International Conference on Machine Learning*.

Figure 3: This figure depicts absolute difference between the actual calorie values and the predictions generated by the XGBoost model. The largest error of 95.5 is between White Chocolate Mocha.