Inferential Statistics

Instacart Market Basket Analysis – Capstone Project I

Raviteja Bodla
Springboard Datascience Career Track

Introduction

As per the visual EDA report, it was observed that items that were added first into the cart have high reorder rate and organic, low-fat and gluten-free foods have high reorder rate. This report presents the results of inferential statistics performed to check the significance of above EDA results.

Added to the cart order and Reorder rate

This test was performed to check if there is any relationship between the added to cart order and reorder rate. To perform the test, we have used Z-test as the reordered variable is a binary variable and added to the cart order is a multivariate feature. There were 6 datasets originally (aisles.csv, departments.csv, orders.csv, order_products_prior.csv, order_products_train.csv and products.csv) which were merged into one dataset 'total1.csv'. From the merged dataset, required columns ('order_id', 'add_to_cart_order', 'reordered') were taken and Z-test was performed at 0.05 significance level.

Results:

1) There is a significant relationship between added to cart order and reorder rate.
2) The p-value obtained was almost equal to 0.

Organic food and Reorder rate

This test was performed to check if there is any relationship between the product name with 'Organic' in it and reorder rate. A new feature (is_organic) was created which has 'True' if 'Organic' was present in the 'product_name' feature. A contingency table was created with reordered as index and is_organic as columns.

is_organic	False	True
reordered		
0	9978458	3885288
1	13184660	6770700

To perform the test, we have used Fisher exact test at 0.05 significance level as the both the features reordered and is_organic are binary features.

Results:

1) There is a significant relationship between the product name with 'Organic' in it and reorder rate.
2) The p-value obtained was almost equal to 0.

Gluten free food and Reorder rate

This test was performed to check if there is any relationship between the product name with 'Gluten' and 'Free' in it and reorder rate. A new feature (is_ glutenfree) was created which has 'True' if 'Gluten' and 'Free' were present in the 'product_name' feature. A contingency table was created with reordered as index and is_ glutenfree as columns.

is_glutenfree	False	True
reordered		
0	13860335	3411
1	19950093	5267

To perform the test, we have used Fisher exact test at 0.05 significance level as the both the features reordered and is_ glutenfree are binary features.

Results:

1) There is a significant relationship between the product name with 'Gluten' and 'Free' in it and reorder rate.
2) The p-value obtained was almost equal to 0.001.

Low fat food and Reorder rate

This test was performed to check if there is any relationship between the product name with 'Low' and 'Fat' in it and reorder rate. A new feature (is_ lowfat) was created which has 'True' if 'Low' and 'Fat' were present in the 'product_name' feature. A contingency table was created with reordered as index and is_ lowfat as columns.

is_lowfat	False	True
reordered		
0	13689203	174543
1	19556758	398602

To perform the test, we have used Fisher exact test at 0.05 significance level as the both the features reordered and is_ lowfat are binary features.

Results:

1) There is a significant relationship between the product name with 'Low' and 'Fat' in it and reorder rate.
2) The p-value obtained was almost equal to 0.