*Predictable Purchases First Draft*

Jordan Deagan

Computer Science

New College of Florida

Sarasota, US

Jordan.deagan17@ncf.edu

***Abstract*—This is the first draft of my final Report in Data Mining. It covers the parts that I will include. The Introduction covers my question and reasonings, Data has my source and data tables I am using, Methodology has the algorithms I will use and the way I will create and gather results, Results will describe how my results will be displayed in the final paper, Conclusions has my potential topics and ideas, and Future Work goes over what I might include there depending on what happens.**

I. INTRODUCTION

*A. Question*

The initially idea for this project was a simple association check for commonly bought items with beef jerky, trail mix, and protein bars. This has been changed to become a predictive model for purchasing certain products. When given the time and day a person makes a purchase, as well as that person’s previous purchases, can it be predicted that a specific item, or type of item, will be in their purchase.

*B. Reasoning*

I am asking this question because I want to see how I can turn associations into predictions, and I want to find if human behavior is predictable when it comes to buying groceries. This question is very important to online stores as a predictive software would allow them to guess when someone is most likely to buy something, giving them a way to know what products to focus advertising and recommendations for, and to who they will work best on, to maximize profit.

II. DATA

*A. Data Tables*

The data I am using comes from Instacart. Instacart has an open source data set that provides csvs of all available products, aisles, and departments in the store, orders made organized by users and purchase order, the contents of orders split by most recent to all prior and organized by the order they were put in the cart. I will not be using the department csv, and the only purpose for the aisle csv is to identify visually what aisle IDs mean, and which ones I want to sort on. The other four csvs are what I am preprocessing on, training on, and testing on.

*B. Preprocessing*

For preprocessing, I will create a list of the product IDs for all instances of beef jerky that isn’t dog food, trail mix that isn’t a cookie or bar, and all instances of protein bars. I will create 3 new data sets containing all purchases from users who bought a certain item. I will also create 3 new sets containing the purchases from these users, but from before they bought that item for the first time. Once I run everything through Apriori, the training data will be created using the associations created for first time buyers of an item, the associations for those who have bought an item before, the day of the week, time of day, and how long it has been since they last bought something. I will split this data set in ten and do training using K-folds method.

III. METHODOLOGY

In the previously purchased data sets I create, I will use Apriori to find associations, and will keep those in separate files. Using K-folds cross validation, I will run the prior purchases data sets, containing references to the products in all prior purchases, time of day, day of the week, and days since last purchase, through a predictive algorithm, which will use all of these and the results of Apriori to create a propensity model of the data. I will test this model on the related data of the most recent purchases for each user. To check if the models are accurate, I will check their output for each purchase against the associated product list, which is hidden from the model, to see if the item is actually in the purchase or not. I will keep track of true positives, false positives, false negatives, and true negatives to determine how accurate the model is both during training and testing.

IV. RESULTS

First I will mention the results of the Apriori training, with some of the associations being displayed. Then I will show the accuracy of the predictive models training and testing, probably a table to explain how I got my accuracy, and the rule or tree I develop as a result. I will go over what the rule or tree means a little bit here as well. All three products will have their own section, with a final section comparing the results and accuracies.

V. CONCLUSIONS

I will first go over whether or not I was able to make an accurate predictor, and whether or not I think it is an appropriate one. Part of the concerns will be the broadness of my products, each of which having a large list of possible items to select, instead of finding a specific product to train and test on. Other concerns might be the algorithms used. If I was unable to get an accurate predictor, I will bring up what I thought was the issues with my process, and what I could have done better in hindsight. If I am successful, I will bring up ways that I think I could have made the process easier, more efficient, or better in general

VI. FUTURE WORKS

As far as dealing with this problem in the future, if I am successful this time, my main concern will be making it more efficient, accurate, or able to deal with more specific or undersampled products. The prospect of also creating the models using a few different algorithms and comparing those results to see which algorithms work better might also be a possible idea for future work. If I fail in making a decent predictor, my future work will mostly involve figuring out what part of my process went wrong and fixing it, as well as seeing if others have made similar models and compare methodologies.

REFERENCES

[1] “The Instacart Online Grocery Shopping Dataset 2017”, Accessed from https://www.instacart.com/datasets/grocery-shopping-2017 on 10/26/18