Michelle Rice

DSC630-T301 Predictive Analytics

Term Project Milestone 2

Predicting Grocery Purchases

Project Proposal

6/19/2021

1. **Introduction**

**1.1 Background**

A key component to retail success is understanding the customers and their buying preferences and patterns. If we know what a customer is likely to want to buy before they come to the store, or shop online, we can use that information to help manage inventory, promotions and targeted marketing. The best way to understand what customers are likely to buy in the future is to look at what they have purchased in the past. Some of this is simple average sales: for example, if we know we generally sell 500 gallons of milk every day, we would plan our inventory to accommodate for that. For this type of planning, in addition to knowing what products a customer purchases, it is valuable to know how frequently the product is purchased and when it is purchased.

There is an additional component to purchase history that can give us even more information. Rather than looking at a product individually, looking at what products customers purchase together can give us even greater insight. In addition to what is purchased and when, it is also helpful to know what other products are purchased at the same time. This is referred to as basket analysis and can guide us even further in sales and inventory planning and provide information needed for successful marketing and promotions.

**1.2 Problem Statement**

Accurately predicting sales and inventory is especially important for grocery stores since they are dealing with perishable items. If a store over predicts sales, they may have to discard products that have expired or gone bad. On the other hand, under predicting sales can leave them in a situation of not having desired products on hand, therefore losing revenue and upsetting and possibly losing customers. Additionally, not placing or pricing items appropriately could cause lost sales. How can we determine what products are frequently purchased together so that we can market them together, plan the most beneficial store placement of products and plan pricing in order to encourage customers to purchase additional products in their basket? Can we predict basket size and/or items by day or time?

**1.3 Scope**

The scope of my analysis will look at several aspects of the sales data in attempt to find trends for individual items and find relationships between items to determine where there is correlation in purchases. I will specifically look at:

* Volume and frequency of sales for specific items
* Average time between purchases
* Items frequently purchased together
* Sales volume and items sold based on day of week

**1.4 Document Overview**

This document includes an introduction to the project, including the problem statement, how it will add business value and the scope of this particular analysis. The preliminary requirements discuss the technical approach, the data that will be used, the required development and plan for building, deploying and testing a model. Next the document will discuss expected results, and execution and management approach including the project plan and potential risks involved.

1. **Preliminary Requirements**

**2.1 Technical Approach**

I will be using both R and Python to load the data, clean and prep the date, perform descriptive statistics and initial graphical analysis in order to identify any outliers or problems with the data, perform any necessary feature reduction and then build a model and perform the analysis. My initial thought was to perform logistic regression to identify correlation, which can be done in both R and Python, however my research has also directed me to using Association Rules, which is a method that I have not previously used but is a common approach to basket analysis.

**2.2 Data sources**

I am currently looking at a dataset from Kaggle that contains approximately 1500 lines of order data, including an order id, day of the week and time of day of the order, days since prior order and the items included in the order. Items in the order are assigned their own column and represented by 0 or 1 to indicate if they were in the order or not..

An example of some rows of data:



* 1. **Analysis**

The first step in my analysis will be to import the data and to run some totals and descriptive statistics on the products. This will allow me to identify features that could potentially be removed, for example if there is a product that is only purchased one or two times in the entire set of data, it may not be significant enough to add value to the analysis. I will do some grouping of the data based on products as well as days of the week and time of day to look for trends and patterns. Creating graphs of the data will allow me to see outliers that again, might be able to be removed, or could require further research.

**2.4 Requirement Development**

I will need to install the apriori library in order to use Association Rules, but all packages that I need should be available in either Python or R. The development needed will be to import the data and to generate statistics on that data, develop graphs and to code the models that will be used for the analysis.

* 1. **Model Deployment**

Once the data has been cleaned and pre-processed and I have installed the apriori package, I can begin to build Association Rules model. This will allow me to use the dataset to generate association rules between products. Association rules take into consideration not only which products are purchased together, but which product is the driver. For example, cereal and milk might frequently be bought together, but it is more likely that cereal drives the purchase of milk rather than the other way around since milk has many uses other than in cereal, but most people don’t eat cereal without milk. I will use a regression model to find relationships between sales and day of the week and/or time of the day.

* 1. **Testing and Evaluation**

The regression model for sales predictions could be tested by creating a test and train dataset and evaluating the accuracy. The Association Rules Model for basket analysis can be measured on three things:

Support – the proportion of transactions that satisfy a rule

Confidence – how accurate a rule is likely to be

Lift – the strength of the rule compared to random guesses

1. **Expected Results**

The expected outcomes of this project are that we can predict likelihood of sales at a total level and possibly product specific based on day of the week and time of day, we can predict how often products are purchased and what products are frequently purchased together. This information can be provided to the business to help drive decisions about marketing, sales promotions, product placement and pricing.

1. **Execution and Management of Project**
   1. **Project Plan**

|  |  |
| --- | --- |
| **Target Date** | **Milestone** |
| 6/20/2021 | Data selection and project proposal |
| 7/5/2021 | Preliminary analysis complete |
| 7/25/2021 | Initial Project Presentation and Status |
| 8/14/2021 | Final paper and presentation |

**4.2 Project Risk**

There are some unknowns in the data that I feel could pose a risk in accuracy. I do not know how this data was collected, if it is a random sample or if it is all orders within a specific timeframe. I also don’t know how long of a time period this data was collected or what time of year the purchases were made. Therefore, I could be missing other drivers in sales such as seasonal or holiday influences.

**References**

Li, Susan (Sept. 24, 2017). A Gentle Introduction on Market Basket Analysis – Association Rules. towards data science. Retrieved from: <https://towardsdatascience.com/a-gentle-introduction-on-market-basket-analysis-association-rules-fa4b986a40ce>

Orders data, Understanding Grocery Shopping Behavor. <https://www.kaggle.com/karthickveerakumar/orders-data>

Klimberg, Ron, McCullough, B. D., Fundamentals of Predictive Analytics with JMP , Second Edition. Retrieved from: https://learning.oreilly.com/library/view/fundamentals-of-predictive/9781629608013/xhtml/Chapter16.xhtml#\_Toc466459272