

Capstone Project: eCommerce & Retail Analytics

Your Client

True North Groceries (TNG) is an online grocery retailer that sells a wide variety of fresh, frozen, and non-perishable groceries and food products through their mobile app and website. The company was founded by Ajmal Habib in 2014 as a digital-only retailer and their sales have quadrupled in each of the last 4 quarters. Like most start-ups TNG has a very lean operation and each of their 18 employees are asked to work in a wide range of problems outside of their core duties. However, recently the management team at TNG has decided to hire a dedicated Data Analytics resource. They believe leveraging data and analytics will be foundational to their long-term success and that it can no longer be something that's delivered by non-analytics professionals at the side of their desks. TMG is interested in hiring your firm to help them build-out the company's analytics practice; however, they would like to try before they buy and have asked for your company to deliver a proof-of-concept to showcase your firm's ability to deliver incremental value through data and analytics.

Your Task

Your tasks is to understand your client's business and operating model, identify where analytics could be applied to make better decisions, and finally deliver a proof-of-concept solution for to a few of business opportunities and challenges they face.

Given you're engagement is initially just a proof-of-concept the scope and volume of data TNG is willing to share is limited. It is your responsibility to figure out how to best leverage the provided data to solve some of the challenges/opportunities your client is facing. This brief will provide a high-level overview of some of the work your client is looking for, however, you will have the opportunity to interview your client, gather and refine their requirements, and even propose solving for a problem outside of what they might be looking for (this is optional).

Dataset

The client will provide you data in the form of several .CSV files that were extracted from their data warehouse which uses a Relational Database Management System (RDBMS). As such, the data will be normalized and each file will represent a table in their RDBMS. The following files are provided:

departments.csv:

- *department_id*: department identifier (21 unique departments)
- *department*: the name of the department (ex: frozen, bakery)

aisles.csv:

- *aisle_id*: unique identifier for an aisle (134 unique aisles)
- *aisle*: the name of the aisle (ex: energy granola bars, specialty cheeses)

products.csv:

- *product_id*: product identifier (~50,000 unique products)
- *product_name*: name of the product
- *aisle_id*: foreign key
- *department_id*: foreign key

orders.csv:

- *order_id*: order identifier
- *user_id*: customer identifier
- *eval_set*: which evaluation set this order belongs in (see SET described below)
- *order_number*: the order sequence number for this user (1 = first, n = nth)
- *order_dow*: the day of the week the order was placed on
- *order_hour_of_day*: the hour of the day the order was placed on
- *days_since_prior*: days since the last order, capped at 30 (with NAs for order_number = 1)

order_products_<SET>.csv:

- *order_id*: foreign key
- *product_id*: foreign key
- *add_to_cart_order*: order in which each product was added to cart
- *reordered*: 1 if this product has been ordered by this user in the past, 0 otherwise

where, <SET> is one of the following three subsets:

- ***order_products_prior.csv*** : orders prior to that users most recent order (~3.2m orders)
- ***order_products_train.csv*** : training data supplied to participants (~131k orders)
- ***order_products_test.csv*** : used to evaluate any machine learning models you build (~75k orders)
 - should not be used in any analysis, except at very end to access model performance

New Capabilities planned for Mobile App and Web Platform

TNG allows customers to purchase their groceries through their mobile app and web-based platform. The product development and engineering team recently held a strategy session and have decided to prioritize the development of various new product features and capabilities, many of which rely on big data analytics and predictive modeling. In particular, these two new platform features are of interest:

Path-to-purchase Personalized Product Recommendations

The idea here is to recommend products to customers as they browse our mobile app and desktop platforms. These recommendations should be ideally a function of what product(s) a customer is currently browsing during their shopping session. For example, if a client just added Heinz Ketchup to their basket, the app might recommend that they purchase Mustard as well.

While the engineering team will build out the front-end functionality and the marketing team will design the creative and strategize the presentment (ex: a banner or small pop-up window), the “intelligence”

behind what recommendation to make to each client as they browse the TNG app or website is what your firm is looking to solve.

“Buy it again” Email

Another important product feature that TNG is looking to add is a “Buy It Again” email. Some analysis conducted by TMG showed that customers with a high number of repeat purchases tend to be significantly more profitable and less likely to attrite than the average customer. As such TMG has devised a strategy to grow this segment through the use of personalized outbound email marketing. For example, if it looks like a customer tends to purchase Laundry Detergent every 3-4 weeks, can we send them a proactive email? Thinking along these, one could also understand which items are frequently re-purchased and the time between purchases to inform decisions as early as possible (without having to wait several weeks to detect that a client tends to purchase Laundry Detergent every 3-4 weeks).

Bonus/Extra Challenge: Recipe Recommender

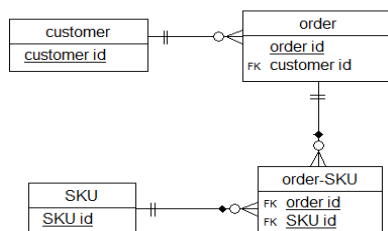
Another idea that has been thrown around by the management team at TMG has revolved around the idea of meal preparation and bundling of ingredients. If you are really looking for a challenge and want to impress your client, feel free to tackle this problem as well. It will most likely require you to merge that dataset provided by the customer with external data assets.

The idea here is to develop a solution that would observe the individual items a client purchases (or has purchased recently) to see if one could make predictions/recommendations for various recipes. For example, if a customer recently purchased eggs, milk, sugar, mascarpone cheese, coffee, and Italian ladyfingers we know they have all/most the ingredients to prepare Tiramisu, a coffee-flavored Italian dessert. If you are looking for even more of a challenge consider a similar but slightly more complex problem where you look at what a client has purchased in a single trip (ex: Mascarpone cheese and Italian ladyfingers) and try to predict/classify what meal/dessert they plan to prepare (note this is not a trivial problem and the solution would have to consider the importance/uniqueness of various ingredients. For example milk, eggs, butter, sugar are used as staples to prepare many meals, while Mascarpone and Italian Ladyfingers are quite unique.)

Your Deliverables

Although this project is open-ended by design, here are some guidelines for what you must deliver:

- **A simple Data Model to illustrate how the various tables (i.e. CSVs) and data sources you’ve used are related through an Entity-Relationship Diagram. See below for an example.**



- **Data exploration, cleansing, and preparation**
 - Even though TNG provided you with the data, there is no guarantee that it's perfect
 - It is your responsible to understand the quality of the data and find opportunities to enhance it (both through data preparation and acquisition of external data where appropriate)
- **Develop an analytical solution to at least two of the following:**
 - Path-to-purchase Personalized Product Recommendations
 - "Buy it again" Email
 - Bonus/Extra Challenge: Recipe Recommender
 - An additional analytical solution to solve a problem/opportunity that is not listed above but you feel will provide value to TNG

This capstone project will be what you work on throughout this course and will be assessed via several independent submissions including a Project Proposal, Sprint 1 Submission, Client Presentation, and Final Report. For more details on deliverables and grade scheme please see the individual rubrics provided for each submission.

References and Additional Notes

TNG is a fictitious client and the material presented in this case study was created by the Course Developer. However, the underlying business problems this capstone looks to tackle are representative of what an online retailer would face. The dataset is based on the data that was originally released by [Instacart](#) with an open-source license. We have used the Instacart dataset as a starting point and made some minor modifications (missing data, errors, outliers, etc.) and arbitrary removal of observations to make solutions to this capstone project unique and prevent plagiarism. The insights, analytics, and models coming out of this project will be unique and distinct from what is publicly available.